ABSTRACT

Title of Dissertation:	USING SOCIAL MEDIA AS A DATA SOURCE IN PUBLIC HEALTH RESEARCH
	Nekabari Sigalo Doctor of Philosophy, 2022
Dissertation directed by:	Associate Professor Vanessa Frias-Martinez College of Information Studies and UMIACS

Researchers have increasingly looked to social media data as a means of measuring population health and well-being in a less intrusive and more scalable manner compared to traditional public health data sources. In this dissertation, I outline three studies that leverage social media as a data source, to answer research questions related to public health and compare traditional public health data sources to social media data sources. In Study #1, I conduct a study with the aim of developing, from geotagged Twitter data, a predictive model for the identification of food deserts in the United States, using the linguistic constructs found in food-related tweets. The results from this study suggest the food-ingestion language found in tweets, such as census-tract level measures of food sentiment and healthiness, are associated with census tract-level food desert status. Additionally, the results suggest that including food ingestion language derived from tweets in classification models that predict food desert status improves model performance when compared to baseline models that only include socio-economic characteristics. In Study #2, I evaluate whether attitudes towards COVID-19 vaccines collected from the Household Pulse Survey can be predicted using attitudes extracted from Twitter. The results reveal that attitudes toward COVID-19 vaccines found in tweets explain 61-72% of the variability in the percentage of HPS respondents that were vaccine hesitant or compliant. The results also reveal significant statistical relationships between perceptions expressed on Twitter and in the survey. In Study #3, I conduct a study to examine whether supplementing COVID-19 vaccine uptake forecast models with the attitudes found in tweets improves over baseline models that only use historical vaccination data. The results of this study reveal that supplementing baseline forecast models with both historical vaccination data and COVID-19 vaccine attitudes found in tweets reduce RMSE by as much as 9%. The studies outlined in this dissertation suggest there is a valuable signal for public health research in Twitter data.

USING SOCIAL MEDIA AS A DATA SOURCE IN PUBLIC HEALTH RESEARCH

by

Nekabari Sigalo

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Vanessa Frias-Martinez, Chair
Professor Beth St. Jean
Professor Kathleen Stewart, Dean's Representative
Professor Quynh Nguyen
Professor Richard Marciano

Acknowledgments

I want to thank my advisor, Dr. Vanessa Frias-Martinez, for being such an amazing advisor and mentor throughout this entire process. I also want to thank my committee members, Dr. Beth St. Jean, Dr. Kathleen Stewart, Dr. Quynh Nguyen, and Dr. Richard Marciano, for their contributions to this work.

I also want to acknowledge all the people in my life who have supported me through this entire journey. To my mom, Veronica Sigalo, for being my personal prayer warrior and for encouraging me to pursue my PhD (in true Nigerian mom fashion!). To my dad, Barida Sigalo, for lending an ear when I needed to vent, and for making everything better with his witty dad jokes. To my siblings – Kasi, Tombari, Barisua, and Jr – for always being there and for believing in me more than I believed in myself. To my amazing niece, Destiny, for being such a light in my life. To my aunts and uncles both in the States and back home in Nigeria – for all their efforts to ensure that they were part of the village that supported me and for encouraging me to keep going. To my best friends, Jere and Chelsea, who have always been there to cheer me on. And to my Heavenly Father, my source, for giving me the strength to get through each day.

Table of Contents

List of Tables

List of Figures

Chapter 1 - Introduction

1.1. Problem Identification

Public health research contributes an immense amount of value to society. With health topics ranging from infectious diseases to food insecurity, public health research provides society with relevant information about disease risk factors and trends, health outcomes and treatments, and public health interventions, and also helps improve the quality of healthcare while reducing healthcare expenditures [1]. Traditional data sources used in public health research include surveys, medical records, claims data, peer-reviewed literature, vital records, surveillance data, and disease registries [2]. While these traditional data sources come with many advantages of their use and have paved the way for a vast body of public health research to be produced, each of these data sources come with their own challenges. For example, while surveys are a great tool for collecting health information to better understand a larger population, surveys are oftentimes expensive to administer, have low response rates, and can be difficult to get detailed information from [3]. Public health surveillance data, which comes from databases and automated electronic reporting systems to monitor disease outbreaks, has proved to be a particularly important data source in controlling the spread of diseases, but one drawback of surveillance data is the sparsity of data from certain geographic areas if responsible agencies do not report it [4]. Medical records are perhaps the most direct and accurate source of health information, but the availability of this information only exists among people who have access to medical care [5] .

Adding to the challenges associated with costly traditional public health data sources is the fact that public health jurisdictions have a history of chronic underfunding and unstable budgets [6]. Federal public health funding is determined annually, making it difficult for public health jurisdictions to plan strategically. A report by TFAH examining federal, state, and local public health funding found that less than 3 percent of the estimated $3.6 trillion allocated annually to health is directed specifically toward public health and prevention in the United States. Considering the pressures on public health funding, it is even more important for public health jurisdictions to find effective and low-cost ways to address the increasing public health challenges of the 21st century.

Over the past few years, alternative data sources for public health research have been adopted by researchers. With social media usage among adults increasing from just 5% in 2005 to 72% in 2021, social media is increasingly becoming a tool for researchers to gain insights into the personal lives of millions of people across the globe [7]. Social media can be broadly defined as a set of online platforms, blogs, and activities that facilitate information sharing, mass communication, crowdsourcing, and collaboration among users [8]. Social media data is self-reported and provides a "snapshot" into the lives of real people, providing researchers with scalable methods of answering targeted research questions. This non-traditional data source has been deemed attractive due the geographic granularity of such novel information, and importantly, the speed of data collection [9].

Social media users are becoming more diverse and more representative of the larger population. Among adults aged 18-29, 84% say they use at least one social media platform; among adults aged 30-49, 81% say they use at least one social media platform; among adults

aged 50-64, 73% say they use at least one social media platform; and 45% of adults 65+ report using at least one social media platform [7]. Research has shown that YouTube & Facebook are the most popular social media platforms, with 81% and 69% of American adults reporting using each platform, respectively [7]. Affinity to specific social media platforms differs by various demographics, such as education, age, and gender [7]. Researchers have increasingly looked to social media data as a means of measuring population health and well-being in a less intrusive and more scalable manner [10]. Social media data have proved useful in evaluating health outcomes in many studies, so it may prove to be a very rich data source for examining other health-related issues, such as food insecurity and COVID-19.

Prior studies have successfully extracted information from social media to address various types of health-related outcomes, relying on the naturalistic observations deduced from social media data to answer questions related to health and well-being [10]. For example, in a study that sought to predict depression among Twitter users, researchers leveraged behavioral cues found in tweets to develop a classifier for depression [11]. In a study that considered Twitter data for various public health applications, researchers conducted syndromic surveillance of serious illnesses, measured behavioral risk factors, and mapped illnesses to various geographic regions [12]. Another study used Twitter to monitor and predict influenza prevalence in the United States by conducting a network analysis of Twitter users and demonstrating the association of social ties and co-location of symptomatic people with one's risk of contracting the flu [13]. A study that sought to develop a publicly available neighborhood-level dataset with indicators related to health behaviors and well-being also

examined the associations between these Twitter-derived indicators and key neighborhood demographics [14].

1.2. Research Objectives

In this dissertation, I present three studies that leverage social media as a data source, to answer research questions related to public health and compare traditional public health data sources to social media data sources.

- *Research Objective #1: Examine the linguistic constructs found in tweets to evaluate the differences in food nutritional value and food consumption behavior of individuals in food deserts compared to non-food deserts.*
    - The following research questions are addressed in this study:
        - RQ 1. Is living in a food desert associated with the food-ingestion language and sentiment of tweets observed among Twitter users?
        - RQ 2. Can the food-ingestion language among Twitter users in a census tract be used to infer census tract-level food desert status?

The first study in this dissertation (Chapter 3) aims to develop, from geotagged Twitter data, a predictive model for the identification of food deserts in the United States, using the linguistic constructs found in food-related tweets. Tweet sentiment and average nutritional value of foods mentioned in tweets were extracted and used to examine the associations between food desert status and the food-ingestion language and sentiment of tweets in a census tract, and to determine whether food-related tweets can be used to infer census tract-level food desert status. The results from this study suggest the food-ingestion language found

in tweets, such as census-tract level measures of food sentiment and healthiness, are associated with census tract-level food desert status. Additionally, the results suggest that including food ingestion language derived from tweets in classification models that predict food desert status improves model performance when compared to baseline models that only include socio-economic characteristics.

The linguistic cues found in tweets may provide some insights into the food environment and food ingestion patterns of individuals. The resulting predictive model can be used as a tool to identify census tracts that may be at risk of becoming food deserts in the future, based on the food conversation found on Twitter, along with other features. Analyzing food-related conversations using tweets presents researchers with the rare opportunity to capture more recent changes in dietary habits and food environment, which would not normally be captured using traditional methods of food environment assessments, such as the USDA's identification of food deserts.

- *Research Objective #2: Examine if aggregate attitudes extracted from COVID-19 vaccine tweets can predict vaccine attitudes reflected in traditional surveys, across metropolitan areas in the United States.*
    - The following research questions are addressed in this study:
        - RQ 1. How do attitudes towards the COVID-19 vaccine found in tweets compare to the attitudes found in surveys?
        - RQ 2. Can aggregate attitudes extracted from COVID-19 vaccine tweets predict vaccine attitudes reflected in traditional surveys?

The second study in this dissertation (Chapter 4) compares social media to a more traditional data source – surveys. In this study, attitudes towards the COVID-19 vaccine found in tweets were compared to the attitudes found in the Census Bureau's Household Pulse Survey. A regression analysis was then conducted to evaluate the ability for sentiments and emotions found in COVID-19 vaccine tweets to predict those expressed in the Census Bureau's Household Pulse Survey. The results revealed that attitudes toward COVID-19 vaccines found in tweets explain 61-72% of the variability in the percentage of HPS respondents that were vaccine hesitant or compliant. The results also revealed significant statistical relationships between perceptions expressed on Twitter and in the survey. The results of this study may suggest that social media data may contain much of the same information found in traditional surveys, with the added benefit of more readily available data and no or low-cost data collection efforts.

- *Research Objective #3: Develop a time series forecasting algorithm that can predict vaccination rates across metropolitan areas using information extracted from tweets.*
    - The following research question will be addressed in this study:
        - RQ 1. Does supplementing forecast models with real-time information found in tweets improve over baseline models that forecast vaccination rates using historical data only?

In the third study (Chapter 5), I examine whether supplementing COVID-19 vaccine uptake forecast models with the attitudes found in COVID-19 vaccine tweets improves over baseline models that only use historical vaccination data. COVID-19 vaccine tweets were used to construct features that were included in the time series forecasting model, such as daily sentiment, emotions expressed in tweets, and user engagement metrics, such as number of

tweets about COVID-19 vaccines, re-tweets, and favorites. The results of this study revealed that supplementing baseline forecast models with both historical vaccination data and COVID-19 vaccine attitudes found in tweets reduce RMSE by as much as 9%. The conversation on social media surrounding COVID-19 vaccines changes daily, as do vaccination rates, so accounting for the vaccine conversation on social media might improve the performance of vaccine forecast models. The results of this study may lead to the development of a predictive tool for vaccination uptake in the United States, which may empower public health researchers and decision makers to design targeted vaccination campaigns in hopes of achieving the vaccination threshold required for the US to reach herd immunity.

The culmination of each of these studies will provide further evidence of the benefits of using social media data for public health research. The overarching contribution of my work will be adapting alternative data sources, machine learning, and natural language processing techniques to assist in public health decision making.

1.3. Organization of the Dissertation

The organization of this dissertation is as follows: in Chapter 2, I present a review of the literature and related work – focusing on natural language processing (NLP) methods, the Twitter API, and key studies that combine NLP methods and Twitter data to conduct public health research. In Chapter 3, I present a study which focuses on predicting food desert status using Twitter data. This study was published in the Journal of Medical Internet Research (JMIR) – Public Health and Surveillance in July 2022. In Chapter 4, I present a study which seeks to validate social media as a data source by predicting the public perceptions of the COVID-19

vaccine in traditional surveys using sentiments and emotions found in tweets. This study is currently under review at Journal of Medical Internet Research (JMIR) – Public Health and Surveillance. In Chapter 5, I present a third study, where I forecast COVID-19 vaccination rates using Twitter data. This study is currently under review at Journal of Medical Internet Research (JMIR) – Infodemiology. In the final chapter, Chapter 6, I present the conclusions, limitations, policy implications, and future directions of this dissertation.

Chapter 2 - Literature Review

2.1. Study #1 - Using Social Media to Predict Food Deserts in the United States

In Chapter 1, I present a study that aims to examine the use of food-ingestion language found in tweets to predict food desert status among census tracts in the United States. Outside of more traditional methods, there is a growing body of work that uses social media as a lens into the food environment and dietary choices of individuals in various geographic locations. The following section provides an overview of relevant studies.

2.1.1. Traditional Food Insecurity Identification Methods

There are several different ways researchers identify and assess food insecure regions and food deserts. A review of the literature determined the most frequently used measures to assess food access are (1) GIS technology/census data, (2) food store assessments, and (3) surveys [15].

2.1.1.1. Food Store Assessments

Researchers have utilized food store assessments to measure food insecurity and identify food deserts in the United States. Food store assessments may include both objective and subjective assessments of the food environment. In a previous study that sought to determine whether or not residents found healthy foods to be accessible in their neighborhoods, objective food store assessments were conducted by simply counting the number of food stores located within a mile of the three Boys and Girls Clubs in the area [16]. Subjective food store assessments were in the form of interviews with study participants, where they were asked open-ended questions pertaining to perceptions and interpretations of

food access, and asked to describe the relationship between social contexts, social conditions, social positions, and access to healthy foods [17]. The study found a disconnect between the objective and subjective food store assessments – food stores identified during the objective assessments were oftentimes not acknowledged by interviewees because of the "less than ideal" quality and nature of the stores [17].

Another study that utilized food store assessments to measure food insecurity involved an instrument that was designed to measure the availability, quality, and preparation of food based on the menus of restaurants [17]. Researchers chose to use the restaurants' menus to reduce bias in the data and avoid interactions between the surveyor and restaurant employees. The instrument also assessed features such as promotions, cleanliness, and service quality [18].

There are several limitations associated with using food store assessments to measure food insecurity and identify food deserts. One possible limitation is sample size. For example, the researchers in [19] recognized the impact of their small, non-random sample size on the ability for study results to be generalized to a larger population. Additionally, researchers often only examine one metro area at a time [19]. Another limitation of using food store assessments for measuring food insecurity and identifying food deserts is the costs associated with conducting these assessments. Studies that use food store assessments require the use of several different data sources to identify food stores, personnel have several different stores to assess, and trained observers must spend a significant amount of time traveling to and from stores and conducting the assessments [20]. Time spent conducting the study is also a limitation associated with this method. In previous work, where the study was conducted over

a long time period, the food environment and access changed rather quickly, leading one to question the relevance of the results [18].

2.1.1.2. GIS Technology and Census Data

Another traditional method for measuring food insecurity and identifying food deserts in the United States incorporates geographic information system (GIS) technology and census data. In this method, researchers use geocoding to map resources, or create density maps that illustrate differences in food security and access in various locations.

One study that used GIS sought to examine the food environment in predominantly Black and low-income areas, suggesting the food environment has an impact on the development of food-related chronic conditions. To assess the food environment in neighborhoods of interest, researchers mapped all the fast-food restaurants in the city of New Orleans, Louisiana and calculated the number of restaurants per square mile in each census tract. Researchers used multiple regression to determine the association between the number of restaurants per square mile and Black and low-income neighborhoods [21].

Another study that used GIS to measure food insecurity sought to examine the availability of supermarkets in neighborhoods of various racial and socioeconomic compositions [16]. Researchers mapped various food stores in Mississippi, North Carolina, Maryland, and Minnesota from the addresses obtained from local health departments, and linked census data to each mapped census tract. Neighborhood wealth was measured by median house values from census data, and racial composition was measured using the proportion of black residents from census data.

There are several limitations associated with using GIS and census data in order to measure food insecurity and identify food deserts. The first limitation to be discussed here is risk of misidentification of food stores in the GIS [16]. GIS maps are pulled from several different data sources, so there will never be perfect accuracy and precision. Another limitation associated with using GIS and census data is changes in food resources over time. For example, there is the possibility that food stores will close over the course of the study, so the mapping and results may be inaccurate [16]. Finally, mapping fails to provide information about food-consumption behavior within food deserts. Although GIS allows us to illustrate things like store density and composition across different regions, this method does not provide insight into what is going on in these areas with inadequate food access, such as food consumption behavior.

2.1.1.3. Surveys

Surveys have also been used to measure food insecurity in the United States. In this method, researchers gather data from randomly sampled households related to household food expenditures, quantities, and consumption for a specified period of time.

[22] used both surveys and census data to identify a link between body mass index and the socioeconomic characteristics of the neighborhoods in which local grocery stores were located. Researchers used data from the Los Angeles Family and Neighborhood Study (LA FANS) database, a collection of 2620 adults from 65 neighborhoods in Los Angeles County from 2000 and 2002. The topics covered in the LA FANS included household income, education, employment, marital status, and residency, as well as information about medical care, food consumption, and entertainment. BMI was calculated using reported height and weight.

Researchers used reported information to identify a link between food-related responses and BMI.

Another study sought to link neighborhood food availability and dietary behavior. Researchers surveyed a sample of participants in the Supplemental Nutrition Assistance Program (SNAP) using the 1996-1997 National Food Stamp Program Survey [22]. One component of the survey included a 1-week food inventory method, in which participant's food consumption was determined during two at-home interviews. Researchers assessed the relationship between observed and reported food consumption and neighborhood food availability using linear regression.

There are several limitations associated with using surveys to assess food insecurity. One limitation is related to the use of existing national surveys to answer questions related to food consumption. Many studies use information from these surveys as a proxy to measure food insecurity. However, pulling food consumption information from existing, more general surveys may not provide sufficient information for measuring food access. For example, [22] acknowledged the fact that there were certain measures that were better at measuring food consumption behavior than the proxies used from the National Food Stamp Program Survey.

Another limitation associated with using surveys to assess food insecurity is self-reporting inaccuracies. Researchers in [22] suggested that self-reported height and weight are often misreported, and the accuracy of these measures varied significantly among different ethnic groups. In a study that specifically used self-reported weight and height to calculate BMI, these inaccuracies and variations likely caused bias and incorrect associations in the results.

2.1.2. Using Social Media to Examine Food Insecurity

Social media is increasingly becoming a source for measuring food insecurity in the United States. In a study that leveraged social media to better understand the health behaviors and well-being of communities, researchers extracted and labeled foods mentioned in geotagged tweets according to their caloric density and "healthiness" [23]. Researchers found tweets with references related to healthy foods were less frequent in low-income census tracts and neighborhoods with higher proportions of minorities. In a study conducted by [24], researchers developed an algorithm to extract nutritional information, such as calorific content, from food-related Instagram posts. These researchers utilized this framework in another study to understand dietary choices and nutritional challenges in food deserts [10]. Researchers analyzed a sample of 3 million geotagged, food-related posts shared on Instagram, and found that food posts originating from food desert census tracts were higher in fat, cholesterol and sugar compared to food posts originating in non-food desert census tracts. Researchers also used the food ingestion language derived from Instagram posts to develop an algorithm for predicting census tract food desert status, with accuracy over 80%.

In a study by [25], researchers examined dietary preferences across various geographic regions using recipes accessed on the web as a proxy to measure food consumption. In this study, researchers found qualitative agreement between the measured sodium content in accessed recipes over time and time series of hospital admissions rates for congestive heart failure in a hospital in Washington, D.C.

To examine the dietary choices of Americans, [26] analyzed tweets from 210,000 Twitter users, linking their dining experiences to their social networks and demographics. By extracting

the caloric density of foods mentioned in tweets, researchers found correlations between the caloric values of the foods mentioned in tweets and state-wide obesity rates. Researchers also developed a model to predict county-wide obesity and diabetes statistics using the nutritional content of the foods mentioned in tweets and the demographic variables of the respective geographic regions.

2.2. Study #2 - Validating social media as a data source: Public perceptions about COVID-19 vaccines in tweets compared to traditional surveys

In Chapter 2, I present a study that aims to predict the public perceptions of COVID-19 vaccines in traditional surveys using the sentiments and emotions found in tweets. Several researchers have already examined public perceptions of COVID-19 vaccines, but this research has been conducted using a variety of different data sources. For example, some studies examined COVID-19 vaccine hesitancy and acceptance using surveys, while others extracted information from social media platforms. The following section provides an overview of these relevant studies.

2.2.1. Examining COVID-19 Vaccine Perceptions using Survey Data

As previously stated, many studies leveraged surveys to examine COVID-19 vaccine hesitancy and compliance. In April 2020, [27] administered four online, nationally representative surveys to adults in France in order to identify reasons why individuals would or would not take the COVID-19 vaccine once one became available. Researchers found that nearly a quarter of respondents refused to take the COVID-19 vaccine once one was made available to them, citing reasons such as not trusting vaccines in general, concerns about the

expedited vaccine development process, and lack of fear of COVID-19, deeming the vaccine unnecessary.

A cross-sectional, self-administered survey was conducted in [28] to evaluate COVID-19 vaccine intent among nurses in Hong Kong, China. Researchers found higher rates of vaccine hesitancy compared to vaccine acceptance, with nurses citing concerns about the safety and efficacy of the vaccines. An online survey conducted in [29] attempted to identify the predictors of intent to vaccinate against COVID-19 among Americans. Nearly 40% of survey respondents refused to vaccinate against COVID-19. Among survey respondents, male, older, white, married, and higher SES individuals were more likely to be vaccinated against COVID-19. Researchers also found that Republicans and Fox News viewers were less likely to get vaccinated, while individuals who were previously vaccinated for influenza were more likely to be vaccinated for COVID-19.

Another study [30] administered online surveys to adults in the US to measure COVID-19 vaccine acceptance. Researchers used regression models to identify factors correlated with vaccine intent and found that the majority of respondents (69%) were willing to get vaccinated. Individuals were more likely to get vaccinated if their healthcare provider recommended it, if they had moderate or liberal political views, or if they felt they were likely to contract COVID-19 in the future. Researchers in [31] combined interviews and surveys to study COVID-19 vaccine acceptability among parents. Most participants stated both them and their children would get vaccinated for COVID-19, citing self-protection as the primary motivation for vaccination. However, participants did cite concerns regarding vaccine safety and efficacy.

2.2.2. Examining COVID-19 Vaccine Perceptions using Social Media Data

There is no shortage of studies that use social media to study vaccine hesitancy and acceptance. With the onset of the pandemic in early 2020, researchers leveraged NLP methods such as sentiment analysis, emotion analysis, and topic modeling to examine vaccine-related perceptions. [32]–[34] collected tweets over the course of the pandemic in order to examine public sentiments and opinions towards COVID-19 vaccines. Researchers in [35] found most tweets to have positive sentiment, but there was also lots of discussion about vaccine hesitancy and rejection. Researchers also found Twitter bots and political activists to be the main culprits for spreading anti-vaccine views online. A study conducted by [36] found mostly negative perceptions of the COVID-19 vaccine among Twitter users.

A topic modeling and emotion analysis conducted by [37] revealed vaccine progress and vaccine instructions as dominant topics over the past year, with overall positive sentiment found in tweets. Trust remained the dominant emotion expressed in tweets. A topic modeling analysis conducted in [38] revealed topics such as vaccine development, vaccine information seeking, financial concerns, vaccine efficacy, and conspiracy theories. Identifying themes and sentiments on social media platforms is beneficial to public health officials in the battle against COVID-19.

2.3. Study #3 - Using COVID-19 vaccine Twitter chatter to predict vaccination rates in the United States

In Chapter 3, I present a study that aims to establish a comparative framework for forecasting COVID-19 vaccination rates in the United States. In this study, I plan to develop a time series model for forecasting vaccination rates using traditional univariate time series methods and compare the performance of this model to a multivariate time series model that

not only accounts for past vaccination rates, but also accounts for the changes in the vaccine-related discussion on Twitter, over time. While there are several studies that use time series methods to forecast COVID-19 cases, very few studies aim to forecast COVID-19 vaccination rates. The following section provides an overview of these relevant studies.

2.3.1. Forecasting COVID-19 Related Measures Using Social Media

There is no shortage of studies that sought to forecast COVID-19 cases using information from social media. Researchers in [39] conducted a study using COVID-19 related terms mentioned in tweets and Google searches to predict COVID-19 waves in the United States. Researchers found that tweets that mentioned COVID-19 symptoms predicted 100% of first waves of COVID-19 days sooner than other data sources. Another study used data from Google searches, tweets, and Wikipedia page views to predict COVID-19 cases and deaths in the United States [40]. Researchers found models that included features from all three sources performed better than baseline models that did not include these features. Researchers also found that Google searches were a leading indicator of the number of cases and deaths across the United States. Another study [41] examined the relationship between daily COVID-19 cases and COVID-19 related tweets and Google Trends. In a study conducted by [42], researchers used reports of symptoms and diagnoses on Weibo, a popular social media platform in China, in order to predict COVID-19 case counts in Mainland China. Researchers found reports of symptoms and diagnoses on the social media platform to be highly predictive of daily case counts. Although each of these studies forecast COVID-19 cases and deaths, none of these studies forecast COVID-19 vaccination rates.

2.3.2. Forecasting Vaccinations

Very few studies have conducted time series forecasting of the COVID-19 vaccinated population in the United States. In a study conducted by [32], researchers developed a time series model to predict the percentage of the US population that would get at least one dose of the COVID-19 vaccine or be fully vaccinated. Researchers projected that by the end of July 2021, 62.44% and 48% of the US population would get at least one dose of the COVID-19 vaccine or be fully vaccinated, respectively. Although this paper also included a separate tweet sentiment analysis, researchers did not include Twitter-related features in the forecast model. Additionally, researchers used aggregated vaccination data for the entire United States, rather than a more granular geographic level.

Another study aimed to evaluate if and when the world would reach a vaccination rate sufficient enough for herd immunity by forecasting the number of people fully vaccinated against COVID-19 in various countries, including the US [43]. In this study, researchers used a common univariate time series forecasting method, Autoregressive Integrated Moving Average (ARIMA), to forecast the future number of fully vaccinated people using only historical vaccination data. Based on the resulting projections, researchers concluded that countries were nowhere near the necessary herd immunity threshold needed to end the COVID-19 pandemic.

A study conducted by [44] sought to predict COVID-19 vaccine uptake using various sociodemographic factors. Although not a time series forecasting model, the results of this study showed that geographic location, education level, and online access were highly predictive of vaccination uptake in the United States. The model predicted vaccine uptake with 62% accuracy.

Although there are very few studies related to COVID-19 vaccination forecasting, other studies have been conducted to predict immunizations for other illnesses. For example, one study analyzed electronic medical records of a cohort of 250,000 individuals over the course of ten years [45]. Researchers developed a model to predict vaccination uptake of individuals in the upcoming influenza season based on previous personal and social behavioral patterns. Another study developed a tool for leveraging immunization related content from Twitter and Google Trends to develop a model for predicting whether a child would receive immunizations [46]. Researchers were able to predict child immunization status with 76% accuracy.

2.4. Twitter Application Programming Interface (API)

With the vast number of studies using Twitter as a data source, it is important to discuss how these data are accessed. The Twitter application programming interface (API) was established in 2006 to make information on Twitter widely available to researchers, companies, developers, and Twitter users [47]. The Twitter API makes it possible for users to develop software that integrates Twitter, as well as provides access to public Twitter data. Academic researchers make up a large proportion of users who take advantage of the Twitter API, using the various Twitter APIs to collect and analyze the public conversations found in tweets [47]. There are three main Twitter access points: the Streaming API, the Search API, and the Historical PowerTrack [48].

2.4.1. Types of Twitter APIs

The Twitter Streaming API is the most common resource for accessing publicly available tweets [48]. The Streaming API is a free resource which allows users to collect tweets as they are posted in real-time but requires end users to maintain an uninterrupted connection to the

server. While the Streaming API is a great resource for collecting tweets as they are posted, there is a limit to the number of tweets that are available to end users [49]. For example, using the Streaming API, users are limited to a 1% random sample of tweets being pushed to the data server at a single point in time. The Twitter Search API is another publicly available, free resource for collecting publicly available tweets. Unlike the Streaming API, the Search API allows users to pull tweets from the past 6 to 9 days, as opposed to pulling tweets in real-time [50]. The final type of Twitter API is the Historical PowerTrack, which provides users with access to historical tweets posted within a specified time frame [51]. Unlike the Streaming and Search APIs, the Historical PowerTrack API is not free and can be very costly [48].

2.4.2. How the Twitter API works

The most common Twitter APIs, such as the Streaming and Search APIs, can be accessed via popular programming languages, such as R and Python. Packages such as tweepy, twarc, and PyTweet in Python and twitteR, streamR, and rtweet in R allow users to connect to the Twitter Streaming API to collect tweets [52]. These packages allow users to establish a connection with the Twitter data server and users can pull tweets based on keywords and/or location. When searching based on location, it is important to remember that tweets may contain two types of geographical metadata: tweet location, which is available when the user enables the sharing of their location at the time the tweet is posted and is tagged with a specific latitude/longitude coordinate, and account location, which is based on the 'home' location that the user provided in their profile [53]. During a location-based search, the Twitter API will first attempt to find location-matched geo-tagged tweets, which are tweets that were sent while the user had their GPS enabled, and therefore have associated latitude and

longitude coordinates [54]. If this does not return matches, or once the matches have been exhausted, the Twitter API will retrieve tweets from Twitter users who have a 'home' location in their public profile that can be reverse geocoded into the location specified in the search query [54].

Consider the following example: if a user would like to collect COVID-19 related tweets in Atlanta, GA, they should establish a connection with the Twitter API using one of the popular Twitter API packages and specify up to 400 keywords related to COVID-19 to filter their search. In order to further filter their search based on the location of interest, Atlanta, GA, the user must specify the location in their search query using the location name (i.e. fetch tweets from Atlanta, GA), location coordinates and radius (i.e. fetch tweets within 10 miles of 33.7490° N, 84.3880° W, the coordinates of the city center), or the location bounding box, which contains the northeast, northwest, southeast, and southwest geocoordinates of any given location, where the Twitter API pulls any tweets that can be reverse geocoded within this bounding box (i.e. fetch tweets that fall within these four geocoordinates). Using the Streaming API, the resulting query will retrieve COVID-19-related tweets geotagged from Atlanta, GA (or associated with Atlanta, GA based on user profile information) in real-time, as they are being posted. Using the Search API, the resulting query will retrieve COVID-19-related tweets geotagged from Atlanta, GA (or associated with Atlanta, GA based on user profile information) from the past 6 to 9 days.

Once a tweet is fetched from the Twitter API, users must extract the desired information from the tweet object. A single tweet object contains several attributes, including the date and time of the tweet, the user name of the account holder, the full tweet text, tweet

location (if available), the number of times the tweet has been liked, quoted, replied to, retweeted, or favorited, and a unique identifier for the tweet [55].

2.5. Natural Language Processing

A key component of each of the three studies in this dissertation is sentiment and emotion analysis. Sentiment analysis is a natural language processing (NLP) method that identifies positive, negative, or neutral sentiment in text documents [56]. Emotion analysis is also an NLP method, but instead of classifying text documents into positive, negative, or neutral buckets, emotion analysis detects human emotions in text, such as fear, anger, anticipation, surprise, trust, sadness, disgust, and joy [57]. These advanced NLP methods are conveniently available for use in computer programming software applications, such as Python and R. Sentiment analysis packages work by comparing the individual words found in text to one or more specified lexicons. Lexicons contain thousands of words classified as positive or negative, and typically have an intensity score associated with each word, in terms of "positiveness" or "negativeness" [58]. Each word is then classified as positive or negative based on the label and score found in the lexicon. The overall sentiment is then calculated based on the matches between the text and the specified lexicon, and the intensities of the matched words [58].

Several studies have employed sentiment or emotion analysis of social media data to answer research questions related to public health. [59] conducted a sentiment analysis of diabetes-related tweets to analyze feelings towards diabetes, as expressed on Twitter. Researchers sought to evaluate the impact of social media on people living with diabetes, to improve public health interventions. In a study conducted by [60], researchers used sentiment analytics to monitor public health concerns in a less-expensive, more scalable manner. Another

study used sentiment analytics to capture patient experience from online posts [61]. Researchers found the sentiments expressed online to be associated with the results of conventional surveys.

The emergent body of work using social media as a data source in public health research has fueled my interest in developing additional use-cases. In this dissertation, I discuss three different scenarios where social media data can be leveraged to address important public health concerns – each having important policy implications.

Chapter 3 - Study #1: Using Social Media to Predict Food Deserts in the United States

3.1. Introduction

3.1.1. Background

Healthy food is vital to everyday life. However, healthy food is not equally accessible to everyone [62]. Food insecurity refers to an individual's lack of sufficient and consistent access to healthy foods that are both affordable and good in quality because of the lack of financial and other resources [63]. In 2018, the United States Department of Agriculture (USDA) estimated that 14.3 million households (11.1%) in the United States were food insecure [63].

Geographic location is one of the most important contributing factors to food insecurity and access to healthy foods [63]. *Food deserts* can be broadly defined as geographic regions where residents do not have sufficient access to fresh fruits, vegetables, and other essential ingredients for healthy eating [64]. Access to healthy foods can be limited because of low availability of grocery stores, low access to sustainable transportation, abundance of perceivably cheaper but unhealthy fast-food options, or a combination of such reasons [65], [66]. Food deserts are prevalent in rural as well as urban regions, implying that regions with an abundance of food options can still be considered food deserts based on the definition of *healthy* food [67].

3.1.2. Identifying Food Deserts

The disparities in healthy food access among underserved communities have fueled the interest of public health practitioners, researchers, and community activists in not only

identifying regions that are currently food deserts but also regions that are at risk for becoming food deserts in the future. The Economic Research Service at the USDA uses various indicators for the official identification of food deserts in the United States at the census tract level. A review of the literature determined that other frequently used measures to assess food access are as follows: (1) geographic information systems (GIS) technology, where researchers use geocoding to map resources and create density maps that illustrate differences in food security and access in various locations [21]; (2) food store assessments, which may include both objective and subjective assessments of the food environment [16]–[20]; and (3) consumer surveys, which allow researchers to gather data from randomly selected households—data regarding household food expenditures and consumption over a specified period [64].

Although each of these food desert identification methods have been widely used and have provided rich insights into food insecurity in the United States, each method comes with unique challenges. For example, GIS technology comes with the risk of misidentification of food stores in the GIS and mapping fails to provide information about food consumption behavior [17]. Food store assessments may be associated with high costs and small, nonrandom sample sizes, as well as significant time spent conducting assessments [20]. Consumer surveys have been found to reflect self-reporting inaccuracies [22]. Each of the challenges to the state-of-the-art approaches present room for another novel approach that uses an alternative, more modern data source. This study examines the use of food ingestion language found on social media, specifically tweets, for predicting food desert status among census tracts in the United States.

3.1.3. Social Media for Public Health Research

Researchers have increasingly looked to social media data as a means of measuring population health and well-being in a less intrusive and more scalable manner [10]. Social media data have proved useful in predicting health outcomes in many studies; therefore, these data may prove to be a very rich source for yet another health-related issue: food insecurity. Using social media data to predict the emergence of food deserts provides a people-centered approach for identifying food deserts by allowing for the examination of the dietary consumption and habits of individuals who reside in food deserts versus those who do not reside in food deserts [68].

Prior studies have successfully extracted information from social media to address various types of health-related outcomes, relying on the naturalistic observations deduced from social media data to answer questions related to health and well-being [11]. For example, in a study that sought to predict depression among Twitter users, researchers leveraged behavioral cues found in tweets to develop a classifier for depression [11]. In a study that considered Twitter data for various public health applications, researchers conducted syndromic surveillance of serious illnesses, measured behavioral risk factors, and mapped illnesses to various geographic regions [12]. Another study used Twitter to monitor and predict influenza prevalence in the United States by conducting a network analysis of Twitter users and demonstrating the association of social ties and colocation of people who were symptomatic with one's risk of contracting influenza [13]. A study that sought to develop a publicly available neighborhood-level data set with indicators related to health behaviors and well-being also examined the associations between these Twitter-derived indicators and key neighborhood

demographics [14]. Another study examined Instagram posts to understand dietary choices and nutritional challenges in food deserts [64]. The study by Gore et al [69] examined the relationship between the obesity rate in urban areas and the expressions of happiness, diet, and physical activity in tweets.

As seen in this study, several other studies similarly leveraged natural language processing methods such as sentiment analysis, emotion analysis, and topic modeling to use social media to answer public health research questions. For example, some studies [32]–[36] collected tweets over the course of the COVID-19 pandemic to examine public sentiments and opinions regarding COVID-19 vaccines. Researchers [37], [38] conducted topic modeling and emotion analyses to identify the themes and emotions related to the COVID-19 vaccines to aid public health officials in the battle against COVID-19.

3.1.4. Study Overview

In this study, I leveraged the linguistic constructs in food-related tweets to develop a classification model for food deserts in the United States. I considered both tweet sentiment and overall nutritional values of foods found in tweets to identify associations between living in a food desert and food consumption.

To our knowledge, this is the first study to develop a model for inferring food desert status among census tracts in the United States using Twitter data. The main objective of this study was to examine the linguistic constructs found in food-related tweets to evaluate the differences in food nutritional value and food consumption behavior of individuals in food deserts versus those in non–food deserts. Our key hypotheses are as follows: (1) living in a food desert is associated with positive mentions of unhealthy foods, such as tweets that mention

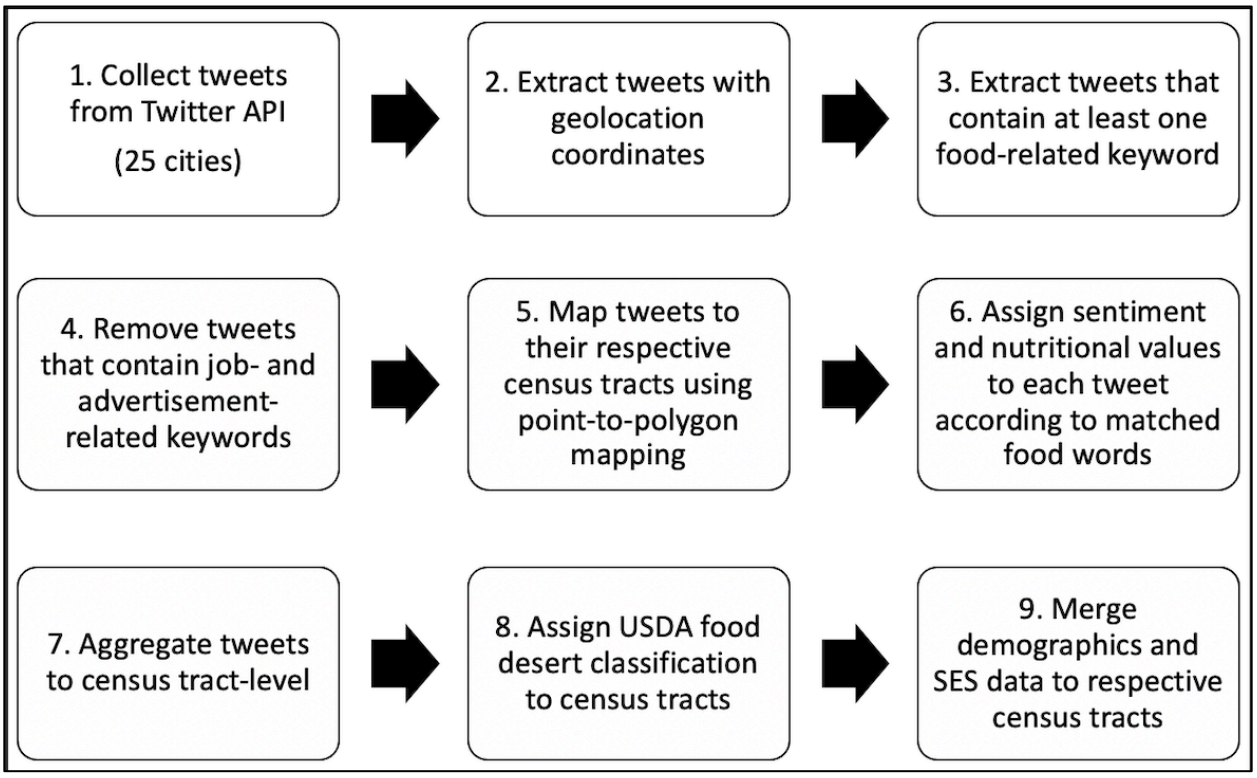
foods that are high in caloric content or low in vital nutrients such as fiber and calcium, and (2) food ingestion language among Twitter users in a census tract can be used to infer census tract–level food desert status.

3.2. Methods

3.2.1. Overview

An overview of the entire data collection and preparation process is illustrated in Figure 3.1 and described in the following subsections.

Figure 3.1 Twitter data collection process, March 2020 to December 2020. API: application programming interface; SES: socioeconomic status; USDA: United States Department of Agriculture.



3.2.2. Data Collection

3.2.2.1. Twitter Data

From March 2020 to December 2020, the Twitter streaming application programming interface (API), which provides access to a random sample of 1% of publicly available tweets, was used to collect tweets (including retweets and quoted tweets) from 25 of the most populated cities in the United States (Table 3.1) [70]. The 25 cities included in this analysis are among the top 50 most populated cities in the United States. However, I decided not to go with the most populated cities (such as New York City, Los Angeles, and Houston) because I wanted to understand whether the framework I developed would be beneficial for smaller cities that are not typically the focus of these types of infodemiology studies. Public health resources directed at improving population health are historically limited and can vary from one public health jurisdiction to the next [71]. Heavily populated cities such as New York City, Los Angeles, and Houston likely have an abundance of resources (both financial and personnel) that can be used to conduct food desert identification using more traditional (expensive) methods. Although the framework outlined in this study would also be useful for heavily populated cities, I believe that less populated cities with fewer resources would benefit the most from this type of study.

Table 3.1 Targeted cities for Twitter data collection, March 2020 to December 2020.

- Albuquerque, New Mexico
- Dallas, Texas
- Atlanta, Georgia
- Baltimore, Maryland
- Colorado Springs, Colorado
- Fresno, California
- Kansas City, Missouri
- Las Vegas, Nevada
- Long Beach, California
- Louisville, Kentucky
- Mesa, Arizona
- Miami, Florida
- Milwaukee, Wisconsin
- Minneapolis, Minnesota
- New Orleans, Louisiana
- Oakland, California
- Oklahoma City, Oklahoma
- Omaha, Nebraska
- Portland, Oregon
- Raleigh, North Carolina
- Sacramento, California
- Tucson, Arizona
- Tulsa, Oklahoma
- Virginia Beach, Virginia
- Wichita, Kansas

When a location-based search is specified, the Twitter API extracts tweets tied to a certain location based on two criteria that are not mutually exclusive: (1) the user has their location enabled for all tweets, in which case these tweets will have specific GPS coordinates, or (2) the user has location information in their profile, such as the city and state they live in, in which case all tweets associated with this user will be tied to this location but without specific geocoordinates. In both cases, these location-tagged tweets are eligible for selection by the Twitter API when a location-based search is specified [53].

As this analysis sought to assign individual tweets to their respective census tracts, all tweets in our sample were required to have specific geolocation information (latitude and longitude GPS coordinates). A parsing module was created to filter out tweets without specific geolocation information. Next, to extract tweets related to food ingestion, tweets were further filtered by a list of 1787 food-related words from the USDA FoodData Central Database (examples are presented in Table 3.2) [72]. Names of popular fast-food restaurants extracted from Wikipedia [73] were also included in this list, as was done in the study by Vydiswaran et al [68]. The complete keyword list can be found in Appendix A.

Table 3.2 Examples of food-related keywords.

- Healthy
  - Acai
  - Apple
  - Apricot
  - Avocado
  - Banana
  - Blackberries
  - Blueberries
  - Cantaloupe
  - Cherries
  - Clementine
- Unhealthy
  - Cheesecake
  - Cupcake
  - Donut
  - Pepsi
  - Sprite
  - Sunkist
  - Red velvet cake
  - Chicken McNuggets
  - Double cheeseburger
  - Zinger burger
- Fast-food restaurants
  - Jack in the Box
  - Chick-fil-A
  - Burger King

- Dairy Queen
- Del Taco
- Taco Bell
- Bojangles
- Checkers
- Popeyes
- Whataburger

Tweets related to job postings and advertisements were filtered out by excluding tweets with hashtags and keywords such as "#jobs," "#hiring," and "#ad." For the purposes of this research, I assumed that the tweets in our sample, which, at minimum, contained at least one of 1787 food-related keywords, were related to food consumption, as was done in the study by Nguyen et al [14]. To assess the impact of this assumption, a random sample of 1000 tweets was selected for manual classification as food related or not food related. Among the 1000 tweets in the random sample, 770 (77%) were classified as food related, whereas 230 (23%), although they contained food keywords, were classified as not food related. Tweets that matched to food words but were not related to food consumption included tweets related to, for example, Apple products (e.g., "I went to the Apple Store to purchase an iPhone") and common city nicknames (e.g., New York City, aka "The Big Apple").

3.2.2.2. Twitter-Derived Features

I referred to similar work conducted by Nguyen et al [14] to classify each food item as healthy or unhealthy. The classification of foods as healthy or unhealthy was subjective and conducted by 2 different annotators (University of Maryland, College Park students – Daniela Nganjo and Pauline Comising). Fruits and vegetables were classified as *healthy food items*.

Unhealthy food items included fried foods, fast-food items, and other food items commonly considered to be unhealthy. The following nutritional values, per 100 g, were obtained for each food item in the list using the USDA FoodData Central Database: calories, calcium, carbohydrates, cholesterol, energy, fat, fiber, iron, potassium, protein, fatty acids, sodium, sugar, vitamin A, and vitamin C.

To measure the healthiness of foods mentioned in tweets, the overall nutritional values of the foods mentioned in each tweet were calculated. To calculate the nutritional values of foods mentioned in each tweet, regular expression matching was used to compare the words in each tweet to the items described in the aforementioned food list (Table 3.2). The keyword-matching algorithm first searched the tweet text for matches to food keywords containing multiple words, then searched the tweet text for matches to food keywords with fewer words. Using this method, the tweet text was searched for keywords with 3 words, for example, before searching for keywords with 2 words, and the tweet text was searched for keywords with 2 words, before searching for keywords with 1 word. For example, both “Burger King” and “burger” were included in the food list. Using this keyword-matching algorithm, a tweet was searched for the keyword “Burger King” first to avoid an incorrect match to the keyword “burger” alone. Once this match was made, the keyword “Burger King” was removed from the tweet text and the remaining tweet text was searched for single-word keywords such as “burger.” Next, the respective nutritional values for each matched food word were then calculated for the corresponding tweet. For tweets having >1 match to food names in the food list, the assigned nutritional value was equal to the average of the nutritional values for all matched food items in the tweet.

3.2.2.3. Sentiment Analysis

To capture the attitudes toward foods mentioned in tweets, I conducted a sentiment analysis of all tweets using the bing lexicon from the tidytext package in R [74]. The bing lexicon provides a label of *negative* or *positive* for thousands of words in the English language. To label the overall sentiment of a tweet, positive expression words were assigned a value of 1, negative expression words were assigned a value of –1, and neutral expression words were assigned a value of 0. An overall *sentiment score* was assigned to each tweet by summing the values assigned to all expression words present in a tweet. Tweets with a positive sentiment score were labeled as having overall positive sentiment, tweets with a negative sentiment score were labeled as having an overall negative sentiment, and tweets with a score of 0 were labeled as having overall neutral sentiment. The resulting tweet sentiment assignments were then used to flag the following types of tweets: tweets that mentioned healthy foods with positive sentiment; tweets that mentioned healthy foods with negative sentiment; tweets that mentioned unhealthy foods with positive sentiment; tweets that mentioned unhealthy foods with negative sentiment; tweets that mentioned fast-food restaurants with positive sentiment; and tweets that mentioned fast-food restaurants with negative sentiment. These tweet-level indicators were later aggregated to the census tract level to produce neighborhood-specific features related to the proportion of tweets that expressed positive or negative sentiment toward healthy foods, unhealthy foods, and fast-food restaurants.

3.2.2.4. Mapping Tweets to Census Tracts

As this analysis examined food desert status at the census tract level, for all census tracts in the 25 cities listed in Table 3.1, each tweet was then mapped to its respective census tract using point-to-polygon mapping of the latitude and longitude coordinates of the

geolocated tweet to the bounding box of the respective census tract [75]. Once each tweet was mapped to a census tract, the tweets were aggregated to the census tract level and the average nutritional content per food item mentioned in tweets within each census tract was calculated. Additional census tract–level food-related Twitter-derived features included the following: (1) percentage of all tweets in a census tract that mention the following with either positive or negative sentiment: healthy foods, unhealthy foods, and fast-food restaurants, and (2) average number of healthy food, unhealthy food, and fast-food mentions per tweet. Tweets with neutral sentiment were not excluded from the analysis sample, but I did not consider neutral sentiment as an independent feature. A complete list of food-related census tract–level features derived from Twitter can be found in Table 3.3.

Table 3.3 Twitter-derived food features.

- Percentage of tweets that mention healthy foods with positive sentiment
- Percentage of tweets that mention healthy foods with negative sentiment
- Percentage of tweets that mention unhealthy foods with positive sentiment
- Percentage of tweets that mention unhealthy foods with negative sentiment
- Percentage of tweets that mention fast-food restaurants with positive sentiment
- Percentage of tweets that mention fast-food restaurants with negative sentiment
- Average number of healthy food mentions
- Average number of unhealthy food mentions
- Average number of fast-food mentions
- Average number of calories per food item (per 100 g)
- Average calcium per food item (per 100 g)
- Average carbohydrates per food item (per 100 g)
- Average cholesterol per food item (per 100 g)
- Average energy per food item (per 100 g)
- Average fiber per food item (per 100 g)
- Average iron per food item (per 100 g)
- Average potassium per food item (per 100 g)
- Average fat per food item (per 100 g)
- Average protein per food item (per 100 g)
- Average saturated fatty acids per food item (per 100 g)
- Average sodium per food item (per 100 g)
- Average sugar per food item (per 100 g)
- Average trans fatty acids per food item (per 100 g)
- Average unsaturated fatty acids per food item (per 100 g)
- Average vitamin A per food item (per 100 g)

- Average vitamin C per food item (per 100 g)
- Average number of calories per healthy food item (per 100 g)
- Average number of calories per unhealthy food item (per 100 g)

3.2.2.5. Food Desert Status

Once all data were collected and aggregated to the census tract level, each census tract was classified as a food desert or not a food desert, according to the USDA Food Access Research Atlas classification of low-income and low-access tracts measured at 1 mile for urban areas and 10 miles for rural areas. The USDA classifies low-income tracts using the following criteria: (1) at least 20% of the residents live below the federal poverty level; (2) median family income is, at most, 80% of the median family income for the state in which the census tract lies; or (3) the census tract is in a metropolitan area and the median family income is, at most, 80% of the median family income for the metropolitan area in which the census tract lies [76]. Low-access census tracts are classified by a significant share (≥500 individuals or at least 33%) of individuals in the census tract being far from a supermarket or grocery store [76]. In total, 7.52% (299/3978) of census tracts with geolocated food-related tweets were classified as low-income, low-access food deserts, measured at 1 mile for urban areas and 10 miles for rural areas.

3.2.2.6. Demographics and Socioeconomic Status Features

Demographic and socioeconomic status (SES) characteristics at the census tract level were pulled from the 2019 American Community Survey and merged onto the census tract–level tweets data set. The demographic variables used in this analysis are presented in Table 3.4.

Table 3.4 Census tract–level demographic and socioeconomic status features extracted from the 2019 American Community Survey.

- Percentage White and non-Hispanic
- Percentage Black or African American
- Percentage other race
- Percentage Asian
- Percentage American Indian or Alaska Native
- Percentage owner-occupied housing units
- Percentage of population living below the federal poverty line
- Number of housing units
- Number of households
- Median family income (US $, 2019)
- Median age (years)
- Population

3.2.3. Data Analysis

Analyses were performed using R software (version 3.5.1; The R Foundation for Statistical Computing) and Python (version 3.8).

3.2.3.1. Evaluating the Association Between Living in a Food Desert and Food Ingestion Language on Twitter

To test the hypothesis that living in a food desert is associated with the food ingestion language of Twitter users, adjusted linear regression was conducted using food desert status as the predictor variable and the SES features listed in Table 3.4 as control features to analyze which Twitter-derived features presented in Textbox 3 were statistically significantly different between food deserts and non–food deserts. Each Twitter-derived feature (Table 3.3) was

designated as the outcome variable in individual linear regression models, as specified in the following equation:

$$y_{Twitter} = \beta_0 + \beta_{FD}x_{FD} + \beta_{SES1}x_{SES1} + \ldots + \beta_{SES12}x_{SES12} + Error$$

where $y_{Twitter}$ = Individual Twitter – derived food feature; β_0 = y – intercept (constant); β_{FD}=food desert classification; and β_{SES12}=each of the 12 demographic and socioeconomic features listed in Table 3.4.

Twitter-derived features that were found to have individual, significant associations with food desert status were later used as features in the classification model for predicting food desert status to test the hypothesis that key food ingestion language found in tweets can be used to infer census tract–level food desert status.

3.2.3.2. Predicting Food Desert Status

To test the hypothesis that food ingestion language found in tweets can be used to infer census tract–level food desert status, classification models were developed using the Twitter-derived food-related nutritional features listed in Table 3.3. I developed 5 different classification models with different sets of features that would allow us to determine which models, if any, show improvements over a baseline model (Table 3.5). The first model, which was considered the baseline model, included demographics and SES features previously found to be strong predictors of food desert status in prior studies [77]; the second model included the demographics and SES features from the baseline model, plus the Twitter-derived food-related nutritional features presented in Table 3.3; the third model included the demographics and SES features from the baseline model, plus the tweet sentiment features; the fourth model included all the features (from models 2 and 3 combined); and the fifth model included the

demographics and SES features from the baseline model, plus all Twitter-derived food-related features found to have a statistically significant association with census tract–level food desert status.

Table 3.5 Classification models for predicting food desert status.

Model	Description	Features
1	Demographics and SES[a] only (baseline)	Demographics and SES features (Table 3.4)
2	Demographics and SES+nutritional values	Demographics and SES features (Table 3.4) and Twitter-derived food-related nutritional features (Table 3.3)
3	Demographics and SES+Twitter mentions sentiment	Demographics and SES features (Table 3.4) and sentiment analysis of Twitter mentions features (Table 3.3)
4	Demographics and SES+nutritional values+Twitter mentions sentiment	Demographics and SES features (Table 3.4), Twitter-derived food-related nutritional features (Table 3.3), and sentiment analysis of Twitter mentions features (Table 3.3)
5	Demographics and SES+statistically significant features	Demographics and SES features (Table 3.4) and Twitter-derived food-related features found to have a statistically significant association with census tract–level food desert status

[a]SES: socioeconomic status.

All features were standardized using minimum-maximum normalization, a method that standardizes data by rescaling the range of individual features to (0, 1), as described in the study by Cao et al [78]. The data were divided into a 70:30 training data and testing data split. Each of the models were built using 5-fold cross-validation to keep computation time to a

minimum. Using the *caret* package in R, each model described in Table 1 was run using several different classification methods: adaptive boosting, gradient boosting, logistic regression, and ensemble methods [79]. The ensemble model combined adaptive boosting, gradient boosting, and logistic regression as base methods. Ensemble modeling is a process that aggregates the predictions of many different modeling algorithms and uses the results of the base models as inputs into a logistic regression model. The ensemble performs as a single model, reducing the generalization error of the prediction compared with the base models alone. The results of each classification method, regardless of performance, are presented in this paper.

3.2.4. Ethics Approval

The University of Maryland College Park institutional review board has determined that this project does not meet the definition of human participant research under the purview of the institutional review board according to federal regulations.

3.3. Results

3.3.1. Overview

A total of 60,174 geolocated food-related tweets were collected during the data collection period. Across the 25 cities in our sample, 3978 census tracts had at least one geolocated food-related tweet, with a median of 4 (IQR 8) geolocated food-related tweets per census tract. Long Beach, California, had the largest representation of tweets (17,303/60,174, 28.75%), as well as the largest representation of users (5189/17,978, 28.86%; Table 3.6). Fresno, California, had the smallest representation of tweets (421/60,174, 0.7%), and Wichita, Kansas, had the smallest representation of users (132/17,978, 0.73%). The maximum number of tweets by a single individual was 1277 (from a user in Long Beach, California). On average,

there were 6686 (SD 3629) tweets collected from 3264 (SD 1385) users each month. The remaining tweet and user statistics can be found in Table 3.6.

Table 3.6 Number of tweets (N=60,174) and users (N=17,978) by city.

City	Number of tweets, n (%)	Number of users, n (%)
Albuquerque, New Mexico	839 (1.39)	224 (1.26)
Atlanta, Georgia	4936 (8.2)	1739 (9.67)
Baltimore, Maryland	2521 (4.19)	684 (3.8)
Colorado Springs, Colorado	847 (1.41)	268 (1.49)
Dallas, Texas	2472 (4.11)	782 (4.35)
Fresno, California	421 (0.7)	153 (0.85)
Kansas City, Missouri	1651 (2.74)	532 (2.96)
Las Vegas, Nevada	2336 (3.88)	872 (4.85)
Long Beach, California	17,303 (28.75)	5189 (28.86)
Louisville, Kentucky	1246 (2.07)	406 (2.26)
Mesa, Arizona	1888 (3.14)	616 (3.43)
Miami, Florida	2576 (4.28)	1080 (6.01)
Milwaukee, Wisconsin	1578 (2.62)	388 (2.16)
Minneapolis, Minnesota	1282 (2.13)	471 (2.62)
New Orleans, Louisiana	2144 (3.56)	641 (3.57)
Oakland, California	2601 (4.32)	614 (3.42)
Oklahoma City, Oklahoma	1143 (1.9)	371 (2.06)
Omaha, Nebraska	742 (1.23)	198 (1.1)
Portland, Oregon	5528 (9.19)	928 (5.16)
Raleigh, North Carolina	1588 (2.64)	454 (2.53)
Sacramento, California	1721 (2.86)	565 (3.14)
Tucson, Arizona	794 (1.32)	250 (1.39)
Tulsa, Oklahoma	622 (1.03)	209 (1.16)
Virginia Beach, Virginia	960 (1.6)	212 (1.18)
Wichita, Kansas	435 (0.72)	132 (0.73)

Table 3.7 displays descriptive statistics of the census tract–level Twitter-derived food features. On average, there was a higher percentage of tweets that mentioned healthy foods with positive sentiment (34%) versus negative sentiment (20%), a higher percentage of tweets that mentioned unhealthy foods with positive sentiment (34%) versus negative sentiment

(17%), and a higher percentage of tweets that mentioned fast-food restaurants with positive sentiment (21%) versus negative sentiment (12%).

Table 3.7 Descriptive statistics of Twitter-derived food features from geolocated food-related tweets.

Twitter-derived food features	Values, mean (SD)
Percentage of tweets that mention healthy foods, positive sentiment	33.8 (0.4)
Percentage of tweets that mention healthy foods, negative sentiment	19.8 (0.3)
Percentage of tweets that mention unhealthy foods, positive sentiment	33.5 (0.4)
Percentage of tweets that mention unhealthy foods, negative sentiment	17.1 (0.3)
Percentage of tweets that mention fast-food restaurants, positive sentiment	21.2 (0.3)
Percentage of tweets that mention fast-food restaurants, negative sentiment	11.7 (0.3)
Average number of healthy food mentions	0.2 (0.3)
Average number of unhealthy food mentions	0.4 (0.4)
Average number of fast-food mentions	0.1 (0.3)
Average number of calories per food item (per 100 g)	155.1 (96.3)
Average calcium per food item (per 100 g)	74 (91.3)
Average carbohydrates per food item (per 100 g)	23.2 (10.9)
Average cholesterol per food item (per 100 g)	57.3 (284.4)
Average energy per food item (per 100 g)	285.1 (115.7)
Average fat per food item (per 100 g)	10.4 (6.9)
Average fiber per food item (per 100 g)	1.7 (1.4)
Average iron per food item (per 100 g)	1.7 (8.5)
Average potassium per food item (per 100 g)	194.5 (93)
Average protein per food item (per 100 g)	7 (4.1)
Average saturated fatty acids per food item (per 100 g)	3.6 (2.5)

Average sodium per food item (per 100 g)	524.7 (962.7)
Average sugar per food item (per 100 g)	11.8 (8.3)
Average trans fatty acids per food item (per 100 g)	0.1 (0.2)
Average unsaturated fatty acids per food item (per 100 g)	2.6 (4)
Average vitamin A per food item (per 100 g)	548.8 (734.5)
Average vitamin C per food item (per 100 g)	7.1 (15.8)
Average number of calories per healthy food item (per 100 g)	67.4 (61.5)
Average number of calories per unhealthy food item (per 100 g)	189.8 (125.9)

Table 3.8 displays descriptive statistics of census tract–level demographics and SES features among census tracts represented in this analysis. Across the represented census tracts, 62.7% (10,682,930/17,038,167) of all residents were White and non-Hispanic, 15.6% (2,657,954/17,038,167) were Black or African American, and 8.9% (1,516,397/17,038,167) identified as other race. The median family income across census tracts was approximately US $82,000, and the median age was approximately 37 years.

Table 3.8 Descriptive statistics of census tract–level demographics and socioeconomic status features extracted from the 2019 American Community Survey.

Characteristic	Values, mean (SD)
Percentage White and non-Hispanic	62.7 (23.4)
Percentage Black or African American	15.6 (21.0)
Percentage other race	8.9 (12.3)
Percentage Asian	7.4 (9.2)
Percentage American Indian or Alaska Native	1.0 (1.9)
Percentage owner-occupied housing units	49.3 (24.8)
Percentage of population living below the federal poverty line	16.2 (12.1)
Number of housing units	1788.4 (863.5)
Number of households	1628.0 (799.1)

Median family income (US $, 2019)	82,371.4 (42,680.1)
Median age (years)	37.0 (6.8)
Population	4283.1 (2243.6)

3.3.2. Hypothesis 1: Living in a Food Desert Is Associated with the Food Ingestion Language and Sentiments of Tweets Observed Among Twitter Users

The adjusted linear regression models confirmed this hypothesis, revealing significant associations between food desert status and 5 of the Twitter-derived food characteristics (Table 3.9). The results show that a census tract being classified as a food desert was associated with an increase in the average cholesterol concentration (per 100 g; *P*=.02*)* per food item mentioned in tweets, a decrease in the average potassium concentration (per 100 g) per food item mentioned in tweets (*P*=.01), and an increase in the average number of unhealthy foods mentioned per tweet (*P*=.03). A census tract being classified as a food desert was also associated with an increase in the proportion of tweets that mentioned healthy foods as well as the proportion of tweets that mentioned fast-food restaurants with positive sentiment (*P*=.03 and *P*=.01, respectively).

Table 3.9 Adjusted linear regression model results examining the associations between living in a food desert and food ingestion language of Twitter users.

Twitter-derived food features	β coefficient	*P* value	SE	R-squared
Percentage of tweets that mention healthy foods, positive sentiment	.077	.03	0.036	0.003
Percentage of tweets that mention healthy foods, negative sentiment	.023	.44	0.031	3.45×10^{-5}

Percentage of tweets that mention unhealthy foods, positive sentiment	–.051	.06	0.027	0.001
Percentage of tweets that mention unhealthy foods, negative sentiment	.022	.32	0.022	3.98×10^{-4}
Percentage of tweets that mention fast-food restaurants, positive sentiment	.096	.01	0.039	0.005
Percentage of tweets that mention fast-food restaurants, negative sentiment	.010	.74	0.032	8.88×10^{-5}
Average number of healthy food mentions	–.002	.54	0.003	9.57×10^{-5}
Average number of unhealthy food mentions	.014	.03	0.006	0.001
Average number of fast-food mentions	–.003	.76	0.010	2.45×10^{-5}
Average number of calories per food item (per 100 g)	.005	.58	0.009	7.93×10^{-5}
Average calcium per food item (per 100 g)	–.001	.60	0.002	7.36×10^{-5}
Average carbohydrates per food item (per 100 g)	–.009	.19	0.007	4.46×10^{-4}
Average cholesterol per food item (per 100 g)	.005	.02	0.002	0.001

Average energy per food item (per 100 g)	.004	.60	0.007	7.37×10^{-5}
Average fat per food item (per 100 g)	–.005	.69	0.012	4.27×10^{-5}
Average fiber per food item (per 100 g)	–.014	.10	0.008	7.26×10^{-4}
Average iron per food item (per 100 g)	-6.44×10^{-4}	.56	0.001	9.04×10^{-5}
Average potassium per food item (per 100 g)	–.008	.01	0.003	0.002
Average protein per food item (per 100 g)	–.002	.88	0.010	6.11×10^{-6}
Average saturated fatty acids per food item (per 100 g)	.007	.31	0.007	2.70×10^{-4}
Average sodium per food item (per 100 g)	–.005	.06	0.002	9.13×10^{-4}
Average sugar per food item (per 100 g)	–.005	.35	0.005	2.29×10^{-4}
Average trans fatty acids per food item (per 100 g)	–.002	.79	0.007	1.78×10^{-5}
Average unsaturated fatty acids per food item (per 100 g)	.002	.72	0.006	3.39×10^{-5}
Average vitamin A per food item (per 100 g)	.004	.58	0.007	8.19×10^{-5}
Average vitamin C per food item (per 100 g)	-5.53×10^{-4}	.71	0.002	3.52×10^{-5}

Average number of calories per healthy food item (per 100 g)	9.58×10^{-4}	.95	0.017	1.92×10^{-6}
Average number of calories per unhealthy food item (per 100 g)	.007	.64	0.015	8.42×10^{-5}

Although I did not expect to see an association between living in a food desert and an increase in mentions of healthy foods with positive sentiment, I hypothesize that such an association might reflect aspirational tweets of individuals who long for healthy food that is not present in their neighborhood (for example, the positive sentiment does not reflect food consumption but rather a wish to increase accessibility).

3.3.3. Hypothesis 2: Food Ingestion Language Among Twitter Users in a Census Tract Can Be Used to Infer Census Tract–Level Food Desert Status

To test the hypothesis that food ingestion language found in tweets can be used to infer census tract–level food desert status, I used various machine learning methods to compare the performance of 4 classification models (Table 3.10). In this paper, I evaluated model performance by comparing each model's area under the receiver operating characteristic curve (AUC) metric, which measures how well each model can distinguish a non–food desert census tract from a food desert census tract. I used this metric, instead of accuracy, for evaluating model performance because this metric is better suited to measure model performance on class-imbalanced data [80], as is the case with the imbalanced food desert classification outcome in our sample data (of the 3978 census tracts, 299, 7.52%, were food desert census tracts). Model 3, which included sentiment features related to food mentions, showed an improvement over the baseline model AUC, using the gradient boosting classification method,

by >7%. This was also the best performing model (AUC 0.823). Model 4, which included all Twitter-derived food-related features, showed an improvement over the baseline model AUC, using the logistic regression classification method, of nearly 19%. These results confirm hypothesis 2, suggesting that the best performing models involve the inclusion of Twitter-derived food ingestion language.

Table 3.10 Model performance.

Method and model[a]		AUC[b]
Adaptive boosting		
	1 (baseline)	0.759
	2	0.749
	3	0.738
	4	0.650
	5	0.723
Gradient boosting		
	1 (baseline)	0.766
	2	0.797
	3	0.823
	4	0.777
	5	0.699
Logistic regression		
	1 (baseline)	0.682
	2	0.720
	3	0.777
	4	0.809
	5	0.663

Ensemble method		
	1 (baseline)	0.769
	2	0.771
	3	0.760
	4	0.641
	5	0.740

[a]Model descriptions (refer to Table 1)—1: demographics and socioeconomic status only (baseline); 2: demographics and socioeconomic status+nutritional values; 3: demographics and socioeconomic status+Twitter mentions sentiment; 4: demographics and socioeconomic status+nutritional values+Twitter mentions sentiment; 5: demographics and socioeconomic status+statistically significant features.

[b]AUC: area under the receiver operating characteristic curve.

3.4. Discussion

3.4.1. Principal Findings

In this study, I sought to address two key hypotheses: (1) living in a food desert is associated with positive mentions of unhealthy foods, such as tweets that mention foods that are high in caloric content or low in vital nutrients such as fiber and calcium, and (2) food ingestion language among Twitter users in a census tract can be used to infer census tract–level food desert status. The study found significant associations between living in a food desert and tweeting about unhealthy foods, including foods high in cholesterol content or low in key nutrients such as potassium. I also found that supplementing classification models with features derived from food ingestion language found in tweets, such as positive sentiment toward mentions of healthy foods and fast-food restaurants, improves baseline models that only include demographic and SES features by up to 19%, with AUC scores >0.8.

3.4.2. Study Findings in Context

Assessing and understanding the food environment in neighborhoods is key to addressing the issue of food insecurity in the United States. The USDA conducts the official identification of food deserts in the United States but this assessment is infrequent and the latest assessment from 2015 is outdated. Other methods such as GIS technology, surveys, and food store assessments, although effective, can be costly and time consuming. Although conducting assessments of food stores provides important insights into the food environment, this study suggests that perhaps residents of census tracts unknowingly provide important information regarding the food environment on Twitter through the food ingestion language found in tweets. Using social media data for food insecurity research allows researchers to examine food consumption in various regions, allowing a comparison of how food ingestion conversation differs between areas where residents have sufficient access to healthy foods and areas where residents do not have sufficient access to healthy foods.

The findings of this study contribute to the literature on food insecurity in the United States by examining the potential effects of living in a food desert on food consumption using Twitter-derived food ingestion features as a proxy to examine food consumption. In this study, I found that food desert status is associated with not only the sentiment toward the types of foods mentioned in tweets but also the nutritional content of foods mentioned in tweets. More specifically, a census tract being classified as a food desert was associated with an increase in the average cholesterol concentration and a decrease in the average potassium concentration (per 100 g) per food item mentioned in tweets, as well as an increase in the proportion of tweets that mention unhealthy foods. A census tract classified as a food desert was also

associated with an increase in the proportion of tweets that mentioned healthy foods and fast-food restaurants with positive sentiment. These findings support prior studies that also found associations between neighborhood characteristics, such as food desert status or fast-food density, and the *healthiness* of tweets in a census tract [14]. These findings also echo the findings in the study by Gore et al [69], which revealed that the prevalence of tweets containing terms related to fruit and vegetables was correlated with lower obesity rates in cities.

This study makes further contributions by examining the predictive ability of food ingestion language derived from tweets on census tract food desert status. This builds upon a similar study that used Instagram posts to understand dietary choices and nutritional challenges in food deserts [64]. In this study, I investigated to what extent ingestion language extracted from Instagram posts was able to infer a census tract's food desert status. This study yielded a model with high accuracy (>80%).

Other similar studies that sought to examine food consumption using tweets across various geographic regions suggest that many of the food-related tweets in an area may be an artifact of visitors to the area, not residents. For example, a study conducted by Mitchell et al [81] showed that travel destinations such as Hawaii have an abundance of tweets with food-related terms. Similarly, the World Happiness Report [82] showed that a larger number of food-related words in tweets were used by users who regularly travel large distances, such as tourists. Although these studies suggest that the tweets I collected may have been from residents or from people who were simply visiting an area, in our study, I decided to consider all tweets under the premise that tweets from nonresidents can also reflect their food consumption experiences when they are in that neighborhood, which still provides some

information regarding the local food environment. It is also important to note that because the data collection period for this study occurred during the height of the COVID-19 pandemic (particularly during travel restrictions and quarantine mandates), this might have allowed us to better capture local movement and tweets from actual residents in these areas because people were being encouraged to stay closer to home and not travel to other areas [83].

Developing an algorithm that predicts food deserts by extracting information from tweets allows researchers to monitor food insecurity more frequently than current methods allow. The use of tweets for research related to food insecurity provides researchers with more frequently updated information, thereby addressing the "lag between capturing information about newly opened and recently closed food retail businesses" [64]. This framework also has implications for policy making and advocacy. On the basis of the results presented in this paper, I recommend the use of similar algorithms by public health officials to encourage the allocation of food resources to census tracts that have been identified as food deserts using the algorithm, especially if these neighborhoods are not currently identified as food deserts according to the USDA's classifications. Public health officials may also leverage this framework to advocate for policy interventions that either prevent food deserts from emerging or increase access to healthy foods in neighborhoods identified as food deserts using the algorithm, minimize the impacts of limited food access, support data-driven decision-making, and encourage grocery store chains to expand into neighborhoods based on need rather than potential profit.

3.4.3. Limitations

Although prior research has proved social media to be a rich data source, it does have some limitations. The ability to pull millions of tweets from a single data source is an attractive

characteristic of Twitter data, but a study conducted by Pew Research Center showed that Twitter users are more likely to be younger than the general population (29% of Twitter users are aged 18 to 29 years compared with 21% of the general population in the United States), more highly educated (42% of Twitter users are college graduates compared with 31% of the general population in the United States), have higher incomes (41% of Twitter users earn at least US $75,000 per year compared with 32% of the general population in the United States), and are more likely to consider themselves Democrats (36% of Twitter users consider themselves Democrats compared with 30% of the general population in the United States) [84]. These demographics raise some concerns in terms of bias in study results and suggest limited ability to generalize results to the larger population.

Adding to the lack of representation among Twitter users is the disparity in Twitter activity among Twitter users. The median number of tweets for Twitter users is only 2 tweets per month. Just 10% of Twitter users account for 80% of the tweets across users in the United States [84]. In studies that use Twitter data, this disparity suggests that a large sample of tweets may only reflect, in reality, a much smaller sample of individuals.

Tweets were collected using the Twitter streaming API, which is limited to a random sample of 1% of all tweets sent by Twitter users at any given time. Of this limited sample of tweets, studies have shown that only approximately 1% to 2% of the tweets from the Twitter streaming API include geolocation information [14]. Because of the nature of this study, our analysis required geolocated tweets, significantly reducing the number of tweets allowed in our sample. As a result, I excluded many census tracts in the 25 cities from our sample because of a lack of geolocated tweets that were also food related. In addition, census tracts that did contain

geolocated food-related tweets may have had only a small number of tweets and these tweets may not be representative of the tweets of all Twitter users who reside in a particular census tract. As our analysis is limited to geolocated tweets, there is also the potential for tweets without location information to differ significantly from tweets with geolocation information, which may suggest biased results because of unknown underlying factors.

Despite these limitations, the results of this study confirm both our hypotheses, demonstrating that food ingestion language found in tweets provides a signal that differentiates food deserts from non–food deserts.

3.4.4. Conclusions

The issue of food insecurity is an important public health issue because of the adverse health outcomes and underlying racial and economic disparities that are associated with insufficient access to healthy foods [64]. Social media data have been increasingly used to answer questions related to health and well-being. Prior research has used various data sources for identifying regions classified as food deserts [64], but this study suggests that perhaps the individuals in these regions unknowingly provide their own accounts of food consumption and food insecurity on social media. In this study, I demonstrated that food desert status is associated with food ingestion language found on Twitter and that food ingestion language can be used to predict and assess the food environment in American neighborhoods.

Chapter 4 - Study #2: Validating social media as a data source: Public perceptions about COVID-19 vaccines in tweets compared to traditional surveys

4.1. Introduction

4.1.1. Background

The implementation of successful COVID-19 vaccine rollout is essential for COVID-19 to remain under control globally. Although vaccines are essential in the global battle against COVID-19, vaccine hesitancy continues to be a barrier for effective and consistent vaccine rollout programs. According to the US Census Bureau's Household Pulse Survey, individuals who reported being hesitant about receiving a COVID-19 vaccine cited concerns about side effects, safety, and lack of trust in the vaccine and/or the government [85]. Although the number of vaccine hesitant individuals continues to decline, the fact that vaccine hesitancy still exists interferes with infection control by vaccination.

Vaccine hesitancy has been fueled in part by the spread of vaccine misinformation both in the media and online. In fact, the COVID-19 vaccine discussion became a popular topic among social media users, with many individuals expressing their concerns about taking the vaccine on social media platforms [86]. Amidst the new normal of self-quarantining and lockdown, Twitter quickly emerged as an important means of COVID-19 communications and discussion [87]. This is in part due to the real-time availability of social media messaging, compared to traditional news reporting methods [88]. Twitter users oftentimes take to the platform not only to announce their own experiences and opinions about the pandemic, but they also see Twitter as a source for up-to-date information about the pandemic [89].

The COVID-19 vaccine conversation on social media platforms has been both beneficial and detrimental to vaccination efforts across the world. Although the exact effect of social media on this unprecedented pandemic is difficult to quantify, there has been a constant battle between facts and misinformation, trust and fearmongering, and hope and anger [90]. Research has shown that social media use plays a role in the low acceptance of vaccines [91], [92]. Therefore, studying the public COVID-19 vaccine–related discussion on social media can help researchers better understand attitudes related to the vaccine [91]. Traditionally, surveys are conducted to understand attitudes related to public health, but the information researchers aim to extract from surveys could also potentially be retrieved from social media.

While there are several studies that examine COVID-19 vaccine attitudes through surveys and social media, to my knowledge, there are no studies that evaluate the ability for Twitter data, a newer data source, to predict the attitudes reflected in traditionally collected surveys, such as the Household Pulse Survey (HPS). In recent years, researchers have looked to social media as a data source, citing the availability of more readily available data and no or low-cost data collection efforts [93]. Traditional paper surveys come with high costs to administer, and even though online surveys eliminate costs of postage, paper, printing, and data entry, these newer online survey services may still cost up to thousands of dollars for one survey [94]. Although relatively inexpensive compared to traditional surveys, online surveys are not always cost-effective [94]. Evaluating the ability for information found in traditional surveys to be predicted by information extracted from social media would suggests that researchers may use this more cost-effective data source to provide us with similar rich information often seen in traditional surveys, or, if saving money comes at the cost of losing rich data.

4.1.2. Study Overview

The main objective of this study is to examine if aggregate attitudes extracted from social media can predict vaccine attitudes collected via surveys. I hypothesize that social media data may contain attitudes similar to those found in traditional surveys, with the added benefit of more readily available data and no or low-cost data collection efforts. Specifically, I hypothesize that there is a direct, positive relationship between (1) positive sentiments and emotions towards COVID-19 vaccines found in Twitter data and the Household Pulse Survey, and (2) negative sentiments and emotions towards COVID-19 vaccines found in Twitter data and the Household Pulse Survey.

4.2. Materials and Methods

4.2.1. Data Collection and Preprocessing

4.2.1.1. Household Pulse Survey Data

In April 2020, the U.S. Census Bureau began releasing a bi-weekly, cross-sectional nationally representative survey, the Household Pulse Survey, in an effort to assess the social and economic impacts of the COVID-19 pandemic on American households [95]. The data from this survey is made publicly available in near real-time, with the purpose of informing federal and state response and recovery planning [96]. On January 6, 2021, the US Census Bureau added COVID-19 vaccine-related questions to the Household Pulse Survey with the goal of understanding the factors contributing to vaccine hesitancy and compliance among Americans [97] [Table 4.1]. These questions assessed COVID-19 vaccine receipt, whether respondents received or plan to receive all required doses, intentions to get vaccinated, and reasons why respondents refused to get vaccinated.

Table 4.1 COVID-19 Vaccine-related Household Pulse Survey questions.

Condition	**Question**	**Responses**
Age 18 years+	Have you received a COVID-19 vaccine?	1) Yes 2) No
Answered "Yes" to "Have you received a COVID-19 vaccine?"	Did you receive (or do you plan to receive) all required doses?	1) Yes 2) No
Answered "No" to "Have you received a COVID-19 vaccine?"	Once a vaccine to prevent COVID-19 is available to you, would you…	1) Definitely get a vaccine 2) Probably get a vaccine 3) Be unsure about getting a vaccine 4) Probably NOT get a vaccine 5) Definitely NOT get a vaccine

The measures of vaccine compliance and hesitancy were assessed for each survey wave overall and by the metropolitan areas in Table 4.2. For the purposes of this analysis, individuals who answered “Yes” to “Have you received a COVID-19 vaccine?” were considered *vaccine compliant* and individuals who answered “No” to “Have you received a COVID-19 vaccine?” and answered they would “Probably get a vaccine”, “Be unsure about getting a vaccine”, “Probably NOT get a vaccine”, or “Definitely NOT get a vaccine” once available were considered *vaccine hesitant*.

Table 4.2 Targeted metropolitan areas for data collection, January - May 2021.

New York-Newark-Jersey City, NY-NJ-PA Metro Area
Los Angeles-Long Beach-Anaheim, CA Metro Area
Chicago-Naperville-Elgin, IL-IN-WI Metro Area
Dallas-Fort Worth-Arlington, TX Metro Area
Houston-The Woodlands-Sugar Land, TX Metro Area
Washington-Arlington-Alexandria, DC-VA-MD-WV Metro Area
Miami-Fort Lauderdale-Pompano Beach, FL Metro Area
Philadelphia-Camden-Wilmington, PA-NJ-DE-MD Metro Area
Atlanta-Sandy Springs-Alpharetta, GA Metro Area
Phoenix-Mesa-Chandler, AZ Metro Area
Boston-Cambridge-Newton, MA-NH Metro Area
San Francisco-Oakland-Berkeley, CA Metro Area
Riverside-San Bernardino-Ontario, CA Metro Area
Detroit-Warren-Dearborn, MI Metro Area
Seattle-Tacoma-Bellevue, WA Metro Area

For this study, I used the Household Pulse Survey microdata from Week 22 to Week 30, which were collected between January 6th and May 25th, 2021 [Table 4.3].

Table 4.3 Household Pulse Survey data collection schedule.

Collection Dates	Week
January 6 – January 19, 2021	22
January 20 – February 2, 2021	23
February 3 – February 16, 2021	24
February 17 – March 2, 2021	25
March 3 – March 16, 2021	26
March 17 – March 30, 2021	27
April 14 – April 27, 2021	28
April 28 – May 11, 2021	29
May 12 – May 25, 2021	30

4.2.1.2. Twitter Data

To align with the Household Pulse Survey data collection period outlined in Table 4.3, the Twitter Streaming Application Programming Interface (API), which provides access to a random sample of 1% of publicly available tweets, was used to collect tweets from the metropolitan areas represented in the Household Pulse Survey [Table 4.2] from January 2021 to May 2021. All tweets had "place" information (usually city and state). The place information found in tweets was used to determine the metropolitan area associated with each tweet. Next, to extract tweets related to COVID-19 vaccines, tweets were further filtered by matching variations of vaccine-related keywords, such as *vaccine, pfizer, moderna, johnson & johnson,* and *dose*. The complete keyword list can be found in Appendix B. The tweets sample was further preprocessed to minimize "noise" resulting from tweets that matched our vaccine-related keywords but did not necessarily reflect the thoughts and opinions of individual Twitter

users. For example, companies often promote job postings and advertisements on Twitter using targeted hashtags in hopes of reaching their target audience. To prevent these tweets from adding noise to the sample, tweets related to job postings and advertisements were removed by excluding tweets with hashtags and keywords such as “#jobs”, “#hiring”, and “#ad”.

4.2.1.3. Sentiment and Emotion Analysis of Tweets

To capture the attitudes found in COVID-19 vaccine-related tweets, a sentiment and emotion analysis of all tweets was conducted using the NRC lexicon from the Syuzhet package in R [74]. The NRC lexicon, developed by Saif Mohammad, contains a list of manually labeled English words and their associations with negative and positive sentiments and common human emotions, such as trust, fear, sadness, surprise, and disgust [98]. The Syuzhet package applies the NRC lexicon by independently evaluating and rating each word or expression within a tweet [99]. The *get_nrc_sentiment* function was applied to all tweets to calculate the valence of eight different emotions (fear, joy, anticipation, anger, disgust, sadness, surprise, trust), along with overall positive and negative sentiment. To assess the accuracy of the sentiment classifier, a random sample of 1000 tweets was selected for manual classification by one individual as having positive, negative, or neutral sentiment. Among the 1000 tweets in the random sample, 734 (73.4%) were accurately classified by the automated sentiment classifier. Due to the high volume of tweets in the study sample, tweets were not manually reclassified.

The percentages of the eight emotions, along with the percentage of positive, neutral, and negative sentiments were calculated at the metropolitan level. For the purposes of this analysis, I used the proportion of tweets with positive sentiment and positive emotions towards vaccines as a proxy to capture vaccine compliance among Twitter users, and the proportion of

tweets with negative sentiment and negative emotions towards vaccines were used as a proxy to capture vaccine hesitancy among Twitter users.

In order to evaluate any potential differences in sentiments and emotions between tweets with location information compared to tweets without location information, a random sample of 1000 COVID-19 vaccine tweets without location information was extracted from a publicly available dataset containing COVID-19-related tweets [100]. A sentiment and emotion analysis was also conducted on this random sample. A two-proportions Z-test, which is a statistical hypothesis test used to determine whether two proportions are different from each other, was conducted to determine any significant differences in sentiments between tweets with location information compared to tweets without location information.

4.2.2. Data Analysis

While there may be several factors contributing to an individual's stance on the COVID-19 vaccine, I examined the differences in COVID-19 vaccine stance by geographic location for each data source and compared the results to determine if the two data sources yielded similar results. This analysis examined if vaccine compliance or vaccine hesitancy differs across geographic regions, and if so, whether these differences are seen in both data sources. Pairwise z-tests for proportions were performed to test the null hypotheses of no difference in COVID-19 vaccine stance across metropolitan areas, with a 0.05 significance level ($\alpha = 0.05$). All p-values were adjusted using the Holm method [101]. To determine whether COVID-19 vaccine attitudes on Twitter can predict the COVID-19 vaccine perceptions that are ultimately expressed in the HPS, two linear regression models were constructed using base R (Table 4.4). In Model 1, the predictor variables were each of the five positive Twitter-derived sentiment and emotion

features, and the outcome variable was the proportion of vaccine compliant HPS respondents. In Model 2, the predictor variables were each of the six negative Twitter-derived sentiment and emotion features, and the outcome variable was the proportion of vaccine hesitant HPS respondents. Since anticipation can be perceived as both positive and negative, this emotion was included as a feature in both models.

Table 4.4 Regression models evaluating the relationship between Twitter sentiments/emotions and HPS vaccine hesitancy and compliance.

Model	Features	Outcome
Model 1	% Positive	% vaccine compliant HPS respondents
	% Joy	
	% Surprise	
	% Trust	
	% Anticipation	
Model 2	% Negative	% vaccine hesitant HPS respondents
	% Anger	
	% Disgust	
	% Sadness	
	% Fear	
	% Anticipation	

4.2.3. Ethics Approval

The University of Maryland College Park institutional review board has determined that this project does not meet the definition of human participant research under the purview of the institutional review board according to federal regulations.

4.3. Results

4.3.1. Descriptive statistics

There was a total of 92,453 tweets from 32,645 users across the 14 metropolitan areas in this study [Table 4.5]. The Los Angeles-Long Beach-Anaheim metropolitan area had the largest representation of tweets (21,500/92,453, 23%), while the New York-Newark-New Jersey

metropolitan area had the largest representation of users (18,400/32,645, 56%). The maximum number of tweets by a single individual was 274 (from a user in the New York-Newark-New Jersey metropolitan area). There was a total of 240,242 respondents to the Household Pulse Survey across the 14 metropolitan areas and nine waves in this study, with the majority of respondents in the Washington-Arlington-Alexandria metropolitan area [Table 4.6].

Table 4.5 Number of tweets (N=92,453) and users (N=32,645) by metropolitan area, January – May 2021.

Metropolitan Area	**Number of tweets, n (%)**	**Number of users, n (%)**
Atlanta-Sandy Springs-Alpharetta, GA	4234 (4.58)	1542 (4.72)
Boston-Cambridge-Newton, MA-NH	3019 (3.27)	1298 (3.98)
Chicago-Naperville-Elgin, IL-IN-WI	5821 (6.3)	2561 (7.84)
Dallas-Fort Worth-Arlington, TX	6203 (6.71)	2299 (7.04)
Detroit-Warren-Dearborn-MI	1082 (1.17)	518 (1.59)
Houston-The Woodlands-Sugar Land, TX	5125 (5.54)	2421 (7.42)
Los Angeles-Long Beach-Anaheim, CA	21500 (23.26)	5429 (16.63)
Miami-Fort Lauderdale-Pompano Beach, FL	1954 (2.11)	849 (2.6)
New York-Newark-Jersey City, NY-NJ-PA	18400 (19.9)	7259 (22.24)
Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	3652 (3.95)	1406 (4.31)
Phoenix-Mesa-Chandler, AZ	4778 (5.17)	1573 (4.82)
San Francisco-Oakland-Berkeley, CA	6376 (6.9)	2008 (6.15)
Seattle-Tacoma-Bellevue, WA	3089 (3.34)	1333 (4.08)
Washington-Arlington-Alexandria, DC-VA-MD-WV	7220 (7.81)	2419 (7.41)

Table 4.6 Number of survey respondents (N=240,242) by metropolitan area, January – May 2021.

Metropolitan Area	Number of respondents, n (%)
Atlanta-Sandy Springs-Alpharetta, GA	12611 (5.25)
Boston-Cambridge-Newton, MA-NH	20078 (8.36)
Chicago-Naperville-Elgin, IL-IN-WI	16044 (6.68)
Dallas-Fort Worth-Arlington, TX	15859 (6.6)
Detroit-Warren-Dearborn-MI	12149 (5.06)
Houston-The Woodlands-Sugar Land, TX	14179 (5.9)
Los Angeles-Long Beach-Anaheim, CA	17006 (7.08)
Miami-Fort Lauderdale-Pompano Beach, FL	11641 (4.85)
New York-Newark-Jersey City, NY-NJ-PA	19730 (8.21)
Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	20240 (8.42)
Phoenix-Mesa-Chandler, AZ	14027 (5.84)
San Francisco-Oakland-Berkeley, CA	17787 (7.4)
Seattle-Tacoma-Bellevue, WA	18615 (7.75)
Washington-Arlington-Alexandria, DC-VA-MD-WV	30276 (12.6)

4.3.2. Attitudes towards COVID-19 vaccines in Twitter Data

A sentiment analysis classified most tweets across all metropolitan areas as having positive sentiment [Figure 4.1]. The Washington-Arlington-Alexandria metropolitan area had the largest proportion of tweets with positive sentiment (58.1%), while the Miami-Ft. Lauderdale-Pompano Beach metropolitan area had the lowest proportion of tweets with positive sentiment (50.9%). Tweets with negative sentiment held the smallest proportions across all metropolitan areas. The Los Angeles-Long Beach-Anaheim metropolitan area had the largest proportion of tweets with negative sentiment (16.4%), while the Miami-Ft. Lauderdale-Pompano Beach metropolitan area had the lowest proportion of tweets with negative sentiment (12.9%). Tweets with neutral sentiment represented between ~28% and 36% of tweets across all metropolitan areas. Examples of tweets expressing positive, neutral, and negative sentiments are displayed in Table 4.7.

Figure 4.1 Distribution of sentiments found in COVID-19 vaccine tweets, by metropolitan area, January – May 2021 (N=92,453).

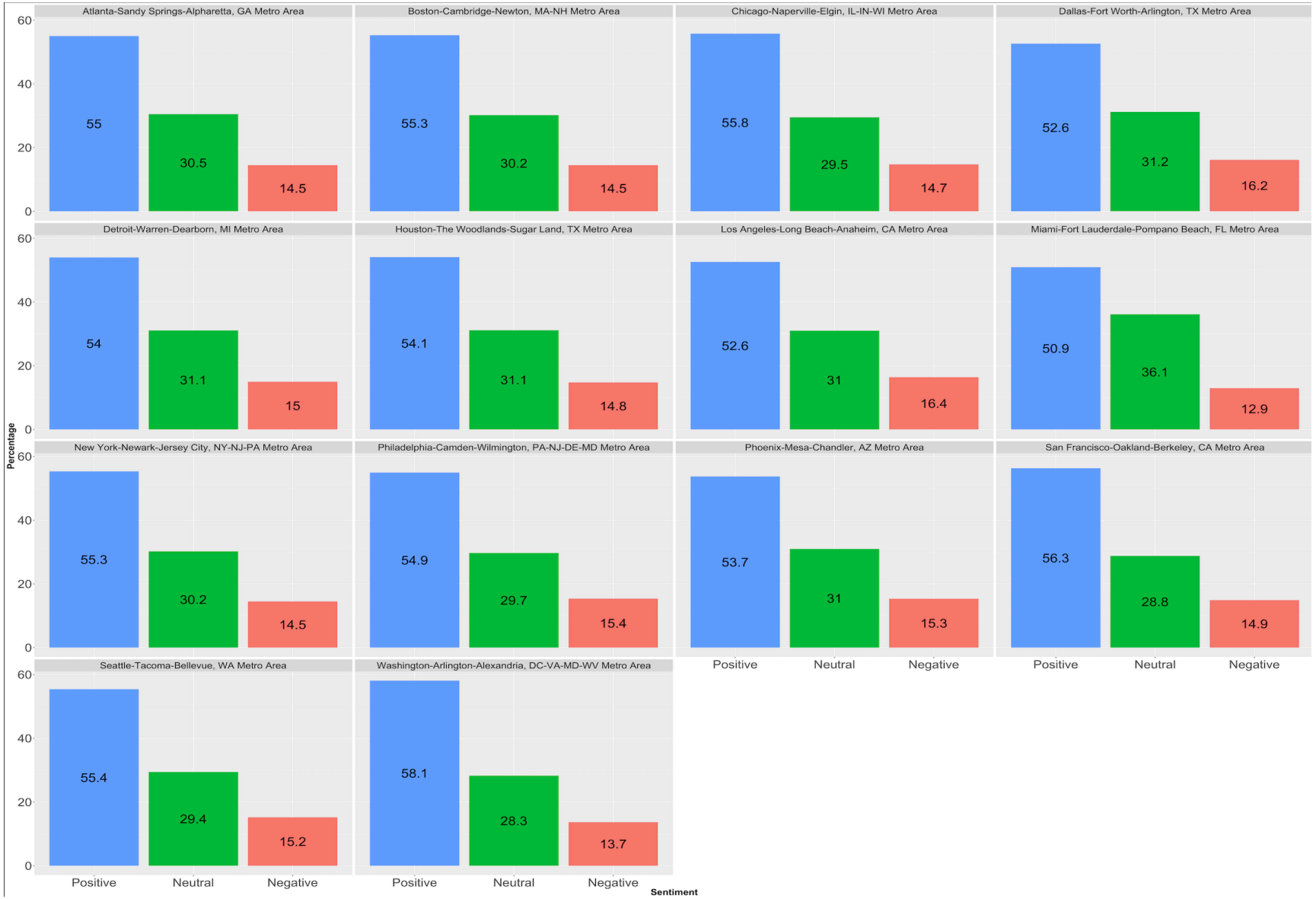


Figure 4.2 Distribution of sentiments found in COVID-19 vaccine tweets without specific geolocation information, January – May 2021 (N=1,000).

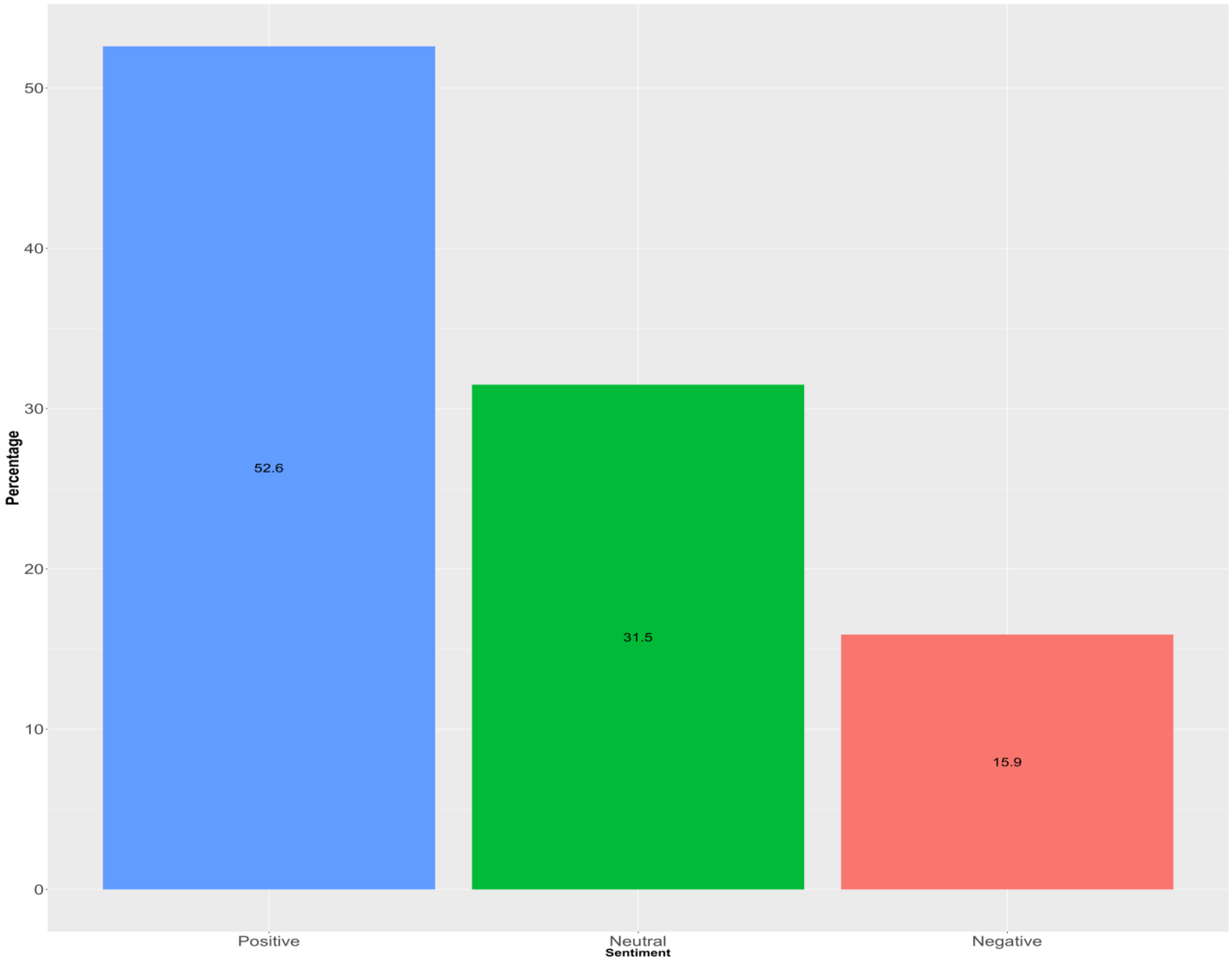


Table 4.7 Examples of tweets expressing positive, negative, and neutral sentiment about COVID-19 vaccines.

Positive sentiments	Neutral Sentiments	Negative Sentiments
Feeling blessed to be healthy this birthday. My two biggest presents are coming in the next week: Inauguration and my second vaccine.	Yeah, I've been vaccinated since early March. Now I wait for everybody else.	This is from the Pfizer v-a-c-c-i-n-e. Please understand these shots cause harm. Injury is REAL & not rare. It's a shame these poor people are being gaslighted, & media giants are censoring them.
Hubby received his first vaccine dose this morning - the sense of relief is for real, folks. #vaccinated	2nd vaccine shot has me like #sleepy #vaccinated #CovidVaccine #gay	they way my people been bugging me about this d*mn vaccine, im not getting that sh*t
With my granddaughter Aurora, Andy, and Elliot. I can see them again and give them a hug now that I am fully Covid 19 vaccinated. I have had both shots plus over two weeks since shot two. Thank you President Biden.	I did my part... It's time to do yours #getvaccinated #Igotvaccinated #nocovid #losangeles #la #california #summer2021 @ North Hollywood, California	No way!! No more lockdowns!! No vaccines!!! Oh and if your so concerned about the virus how about no illegals!!! Thank goodness for New Hampshire and Florida!! Go out
My mom gets her 2nd dose Sunday, big relief!	People are getting vaccinated and meeting up with friends & family IRL again for interesting conversation.	Clearly you are ignorant of the fact that they said even if you get the vaccine you still have to wear a mask, social distance & deal with all the same bull shit draconian orders. Even after blatant evidence you still want to get it. Heres 100% evidence of brain wash mind control
I love so much that I got vaccinated today!	the vaccine can alter your DNA. DNA is double stranded and is in the nucleus, the mRNA used is single stranded and is in the body of the cell, simply encoding your body to make proteins to fight the virus. proceed however you feel comfortable.. but just know.. your DNA is good.	I am 80. You can have my vaccine. I refuse to get one. I take 2 grams of vitamin C hourly. That makes me IMMUNE. Read: Linus Pauling. No mask. I am out every day working & walking in the park. Paul Kangas 4 Governor.
Proud to work for you @bswhealth - my parents received their COVID vaccines this week at BUMC and said it was so quick and easy and the staff were so friendly! Thank you for taking care of them.	It's been two weeks since my second shot. It's for real now... I'm officially fully vaccinated.	F the stupid vaccine

The emotion analysis also revealed trends towards an overall positive sentiment, showing trust as the predominantly expressed emotion in COVID-19 vaccine tweets across all metropolitan areas [Figure 4.3]. The most perceived negative emotions across all metropolitan areas were anticipation and fear. The least perceived positive emotions were joy and surprise, while the least perceived negative emotions were anger and disgust.

In order to evaluate any potential differences in sentiments and emotions between tweets with location information compared to tweets without location information, the proportions of sentiments and emotions in the study sample was compared to the random sample of 1000 COVID-19 vaccine tweets without location information [Figure 4.2 & Figure 4.4]. Similar to tweets with location information, overall sentiment was mostly positive in the random sample of 1000 tweets without location information across the US. In this sample, trust was also the predominantly expressed emotion, with the most perceived negative emotions being anticipation and fear. The least perceived positive emotions were joy and surprise, while

the least perceived negative emotions in this sample were anger and disgust. A two-proportions Z-test confirmed that the proportions of positive, neutral, and negative tweets *with* location information was not significantly different from the proportions of positive, neutral, and negative tweets *without* location information (p-value = 0.2356, 0.4581, and 0.5145, respectively).

Figure 4.3 Distribution of emotions found in COVID-19 vaccine tweets, by metropolitan area, January – May 2021 (N=92,453).

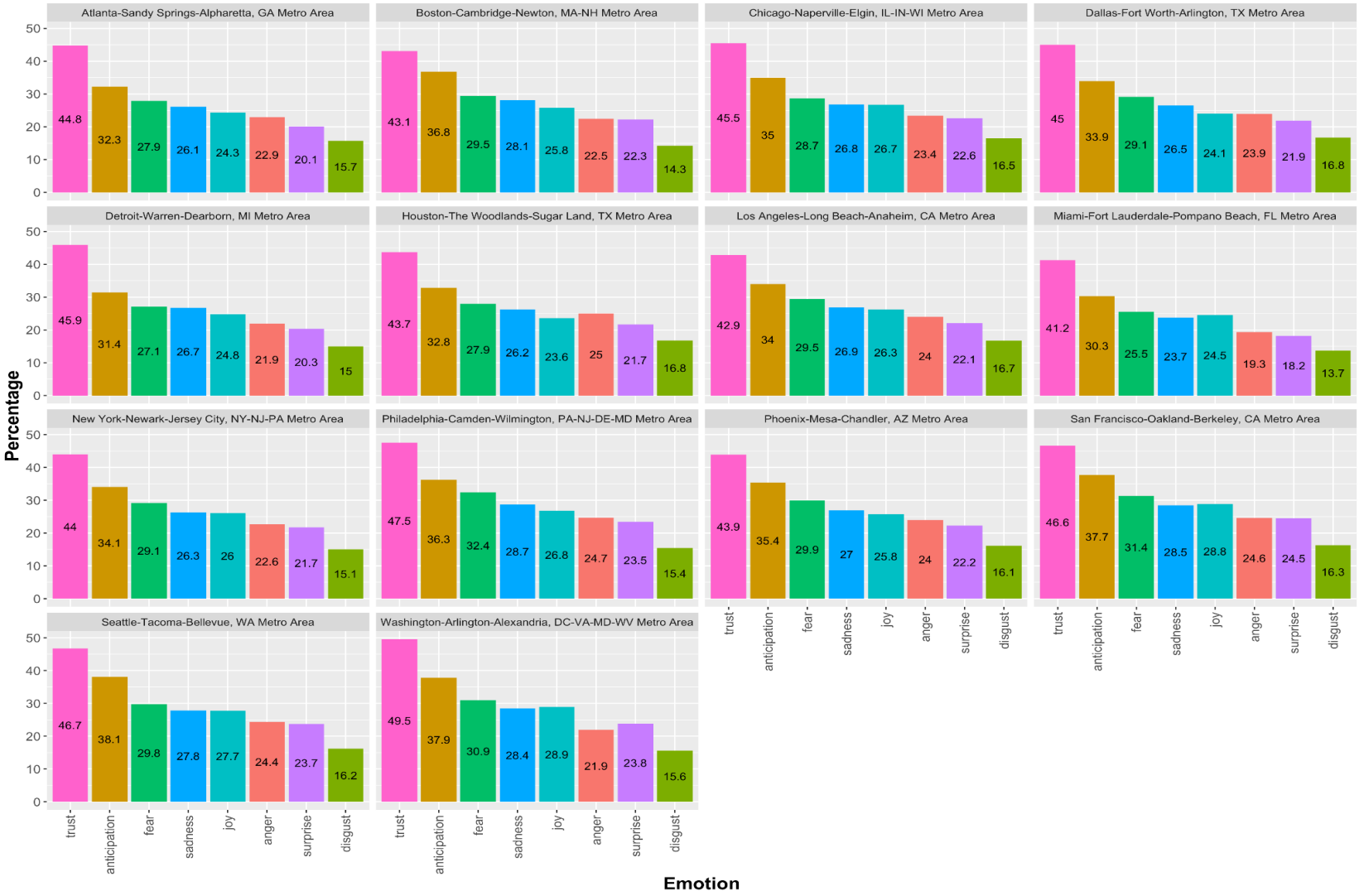
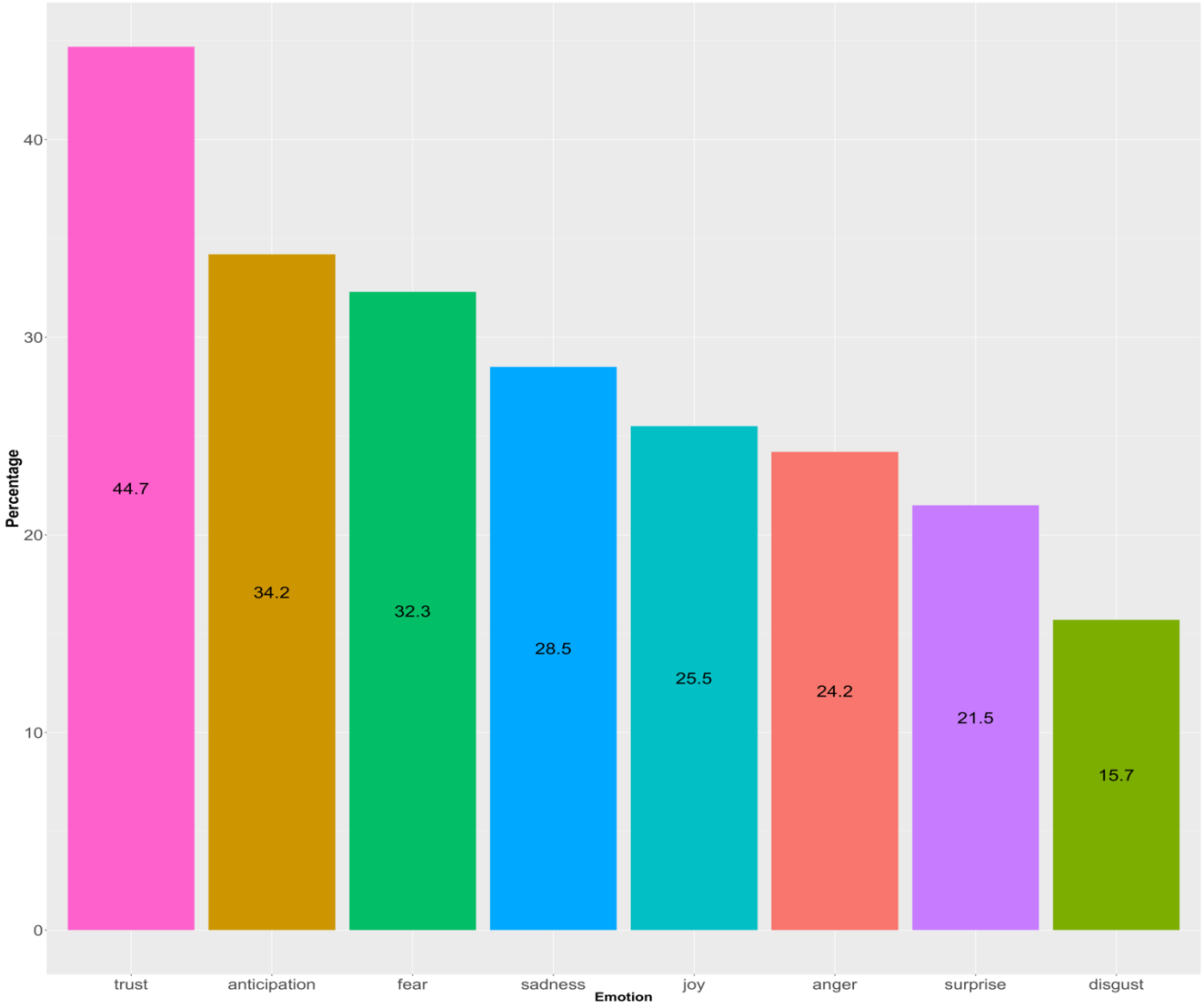


Figure 4.4 Distribution of emotions found in COVID-19 vaccine tweets without specific geolocation information, January – May 2021 (N=1,000).



4.3.3. Attitudes towards COVID-19 vaccines in Household Pulse Survey Data

Most survey respondents across all metropolitan areas indicated they had received a COVID-19 vaccination, ranging from 50.2% of survey respondents in the Phoenix-Mesa-Chandler metro area to 56.4% of survey respondents in the San Francisco-Oakland-Berkeley metro area [Figure 4.5]. Among respondents who indicated they had received a COVID-19 vaccination, the majority also indicated that they received or planned to receive all required doses, except in the Atlanta-Sandy Springs-Alpharetta metropolitan and Phoenix-Mesa-Chandler metropolitan areas (48.4% and 48.3%, respectively) [Figure 4.6]. Among respondents who indicated they had not received a COVID-19 vaccination, the majority indicated that they *probably* or *definitely* would get vaccinated, ranging from 48% of survey respondents in the Phoenix-Mesa-Chandler metro area to 75.2% of survey respondents in the San Francisco-Oakland-Berkeley metro area [Figure 4.7].

Figure 4.5 Distribution of HPS respondents who reported receiving a COVID-19 vaccination, January – May 2021 (N=240,242).

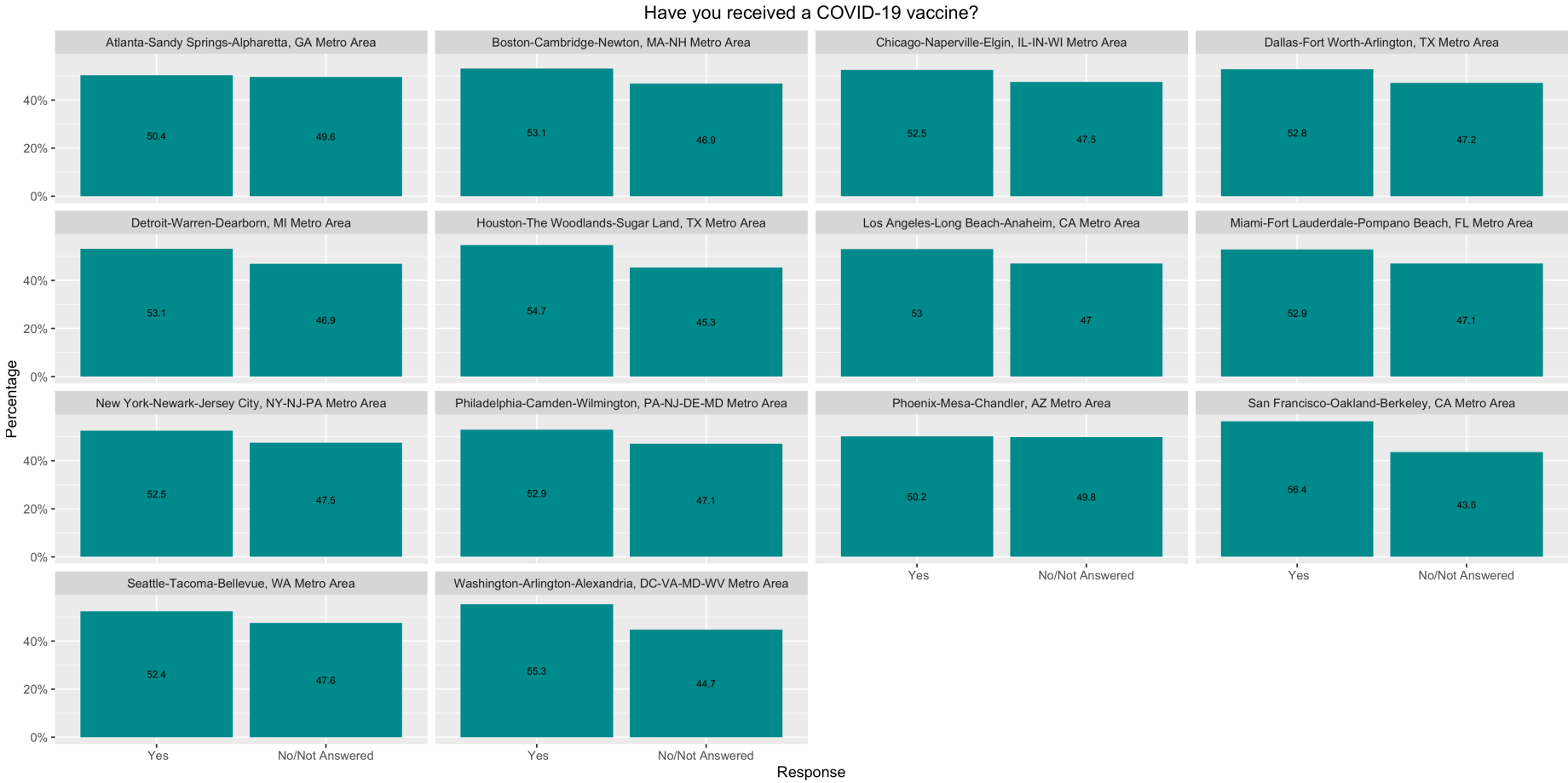


Figure 4.6 Distribution of HPS respondents who reported receiving all required doses of the COVID-19 vaccination, January – May 2021 (N=240,242).

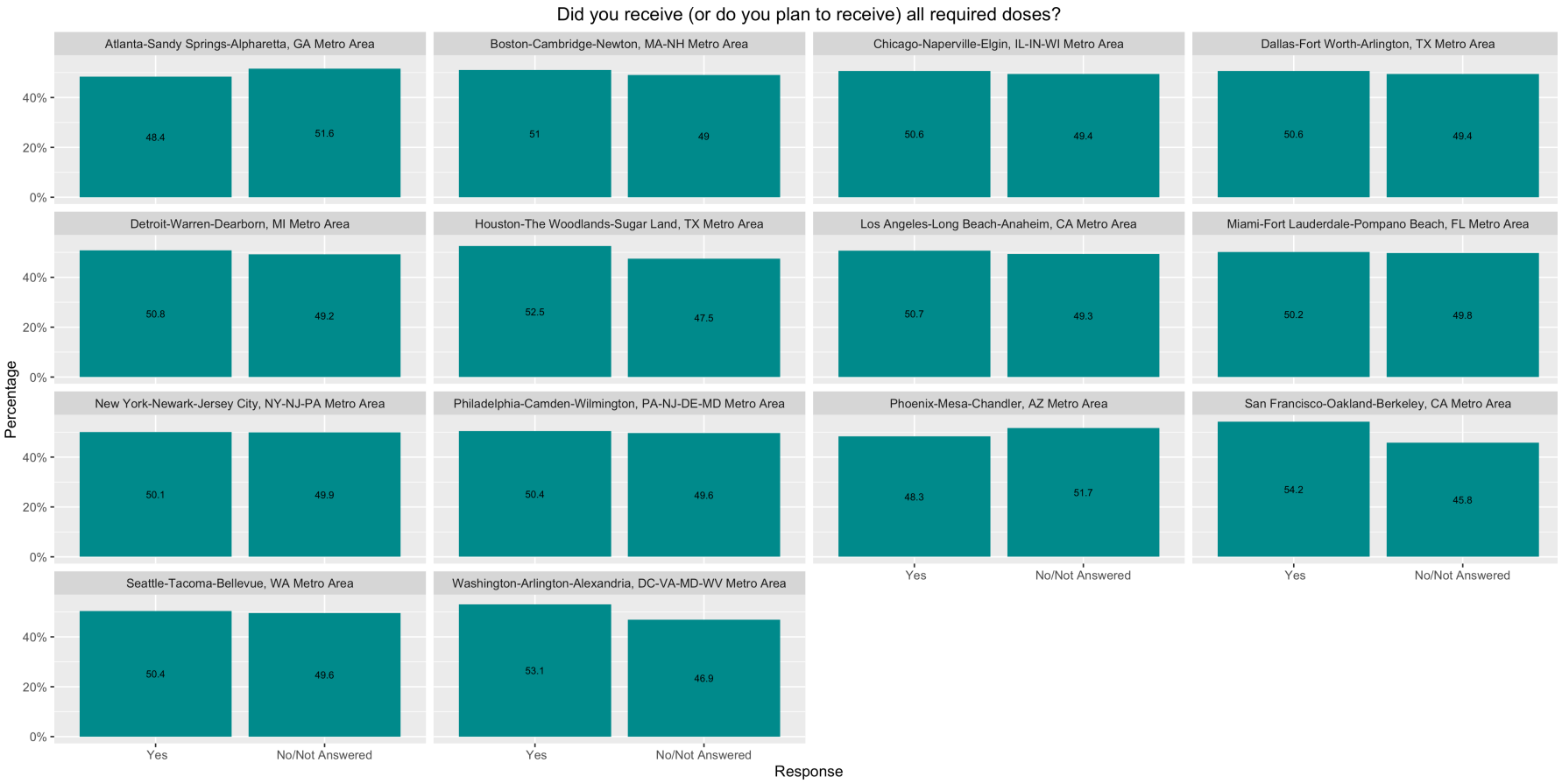
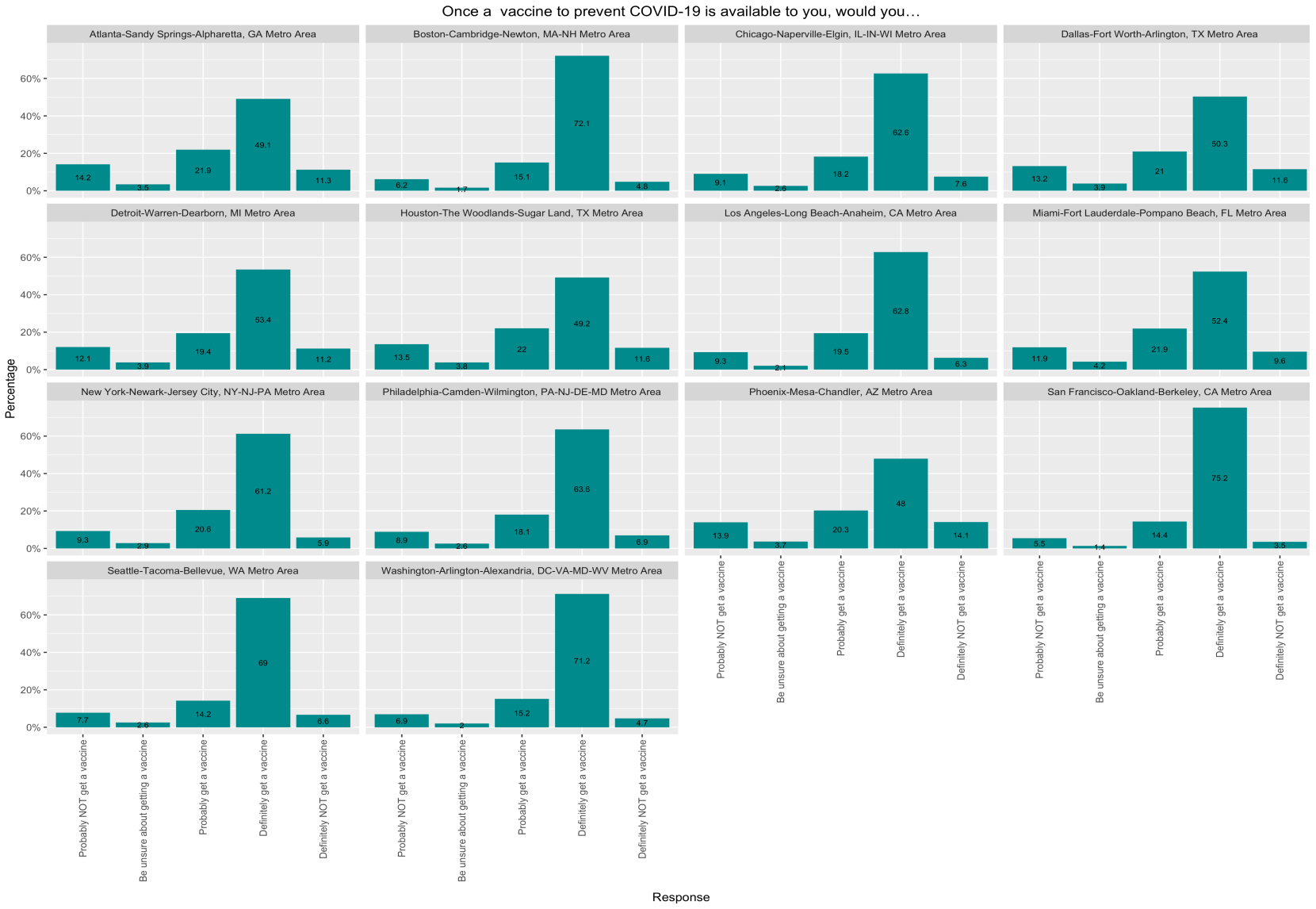


Figure 4.7 Distribution of HPS respondents who reported being vaccine hesitant or vaccine compliant, January – May 2021 (N=240,242).



4.3.4 Public Attitudes towards COVID-19 vaccines: Comparing Twitter data to Household Pulse Survey data

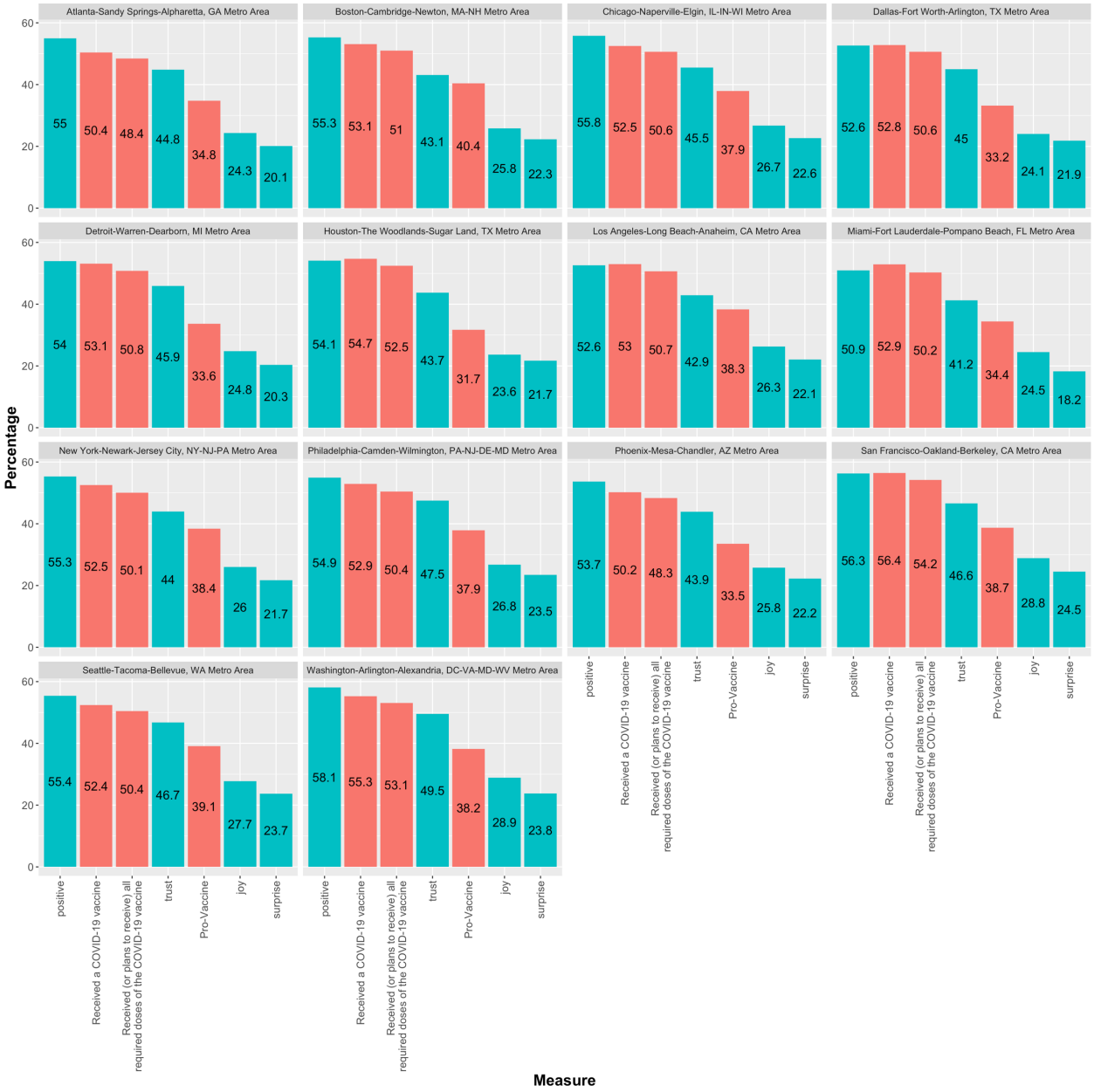
4.3.4.1 Vaccine Compliant Measures

To determine how Twitter data compares to surveys in terms of the public attitudes they reveal towards COVID-19 vaccines, I compared the sentiments and emotions found in tweets to the attitudes towards COVID-19 vaccines expressed in the Household Pulse Survey. Figure 4.8 illustrates the proportion of positive sentiments & positive emotions found in tweets, compared to the proportion of Household Pulse Survey respondents that report being vaccine compliant. In this comparison, I used positive emotions & positive sentiments expressed in tweets as a proxy to measure vaccine compliance among Twitter users. Sentiments & emotions extracted from tweets are shaded blue, while vaccine compliant measures from Household Pulse Survey data are shaded red.

Across all metropolitan areas, the percentage of COVID-19 vaccine tweets expressing positive sentiment was closely aligned with the percentage of survey respondents that indicated they have received a COVID-19 vaccine; received (or plan to receive) all required doses of the COVID-19 vaccine; and “definitely” or “probably” would get a vaccine once a vaccine to prevent COVID-19 was available to them. The San Francisco-Oakland-Berkeley metropolitan area showed the closest alignment between positive COVID-19 vaccine perceptions across the two different data sources, with a difference of 0.1 percentage points between the percentage of COVID-19 vaccine tweets expressing positive sentiment (56.3%) and the percentage of survey respondents that indicated they have received a COVID-19 vaccine (56.4%). The Dallas-Fort Worth-Arlington metropolitan area also showed close alignment

between positive COVID-19 vaccine perceptions across the two different data sources, with a difference of 0.2 percentage points between the percentage of COVID-19 vaccine tweets expressing positive sentiment (52.6%) and the percentage of survey respondents that indicated they have received a COVID-19 vaccine (52.8%).

Figure 4.8 Comparison of vaccine acceptance in Twitter data versus HPS data. Sentiments & emotions extracted from COVID-19 vaccine tweets are shaded blue, while measures from HPS data are shaded red.



The percentage of COVID-19 vaccine tweets expressing trust was also closely aligned with the percentage of survey respondents that indicated they received (or plan to receive) all required doses of the COVID-19 vaccine, or "definitely" or "probably" would get a vaccine once a vaccine to prevent COVID-19 was available to them. The Philadelphia-Camden-Wilmington metropolitan area showed a difference of only 2.9 percentage points between the percentage of COVID-19 vaccine tweets expressing trust (47.5%) and the percentage of survey respondents that indicated they received (or plan to receive) all required doses of the COVID-19 vaccine (50.4%). The Boston-Cambridge-Newton metropolitan area showed a difference of only 2.7 percentage points between the percentage of COVID-19 vaccine tweets expressing trust (43.1%) and the percentage of survey respondents that indicated they "definitely" or "probably" would get a vaccine once a vaccine to prevent COVID-19 was available to them (40.4%). Other positive emotions, such as joy and surprise, did not appear to align with any measures of COVID-19 vaccine compliance in the Household Pulse Survey data.

The results of the proportions tests revealed alignment between the significant differences in the Twitter measure of vaccine compliance (positive sentiment) and the percent of Household Pulse Survey respondents that indicated that they received a COVID-19 vaccine, across the same metropolitan areas, in some cases. For example, both the proportion of Household Pulse Survey respondents that indicated that they received a COVID-19 vaccine and the proportion of tweets with positive sentiment were significantly higher in the San Francisco-Oakland-Berkeley, CA metropolitan area (56.4% and 56.3%, respectively) compared to the Dallas-Fort Worth-Arlington, TX (52.8% and 52.6%, respectively), Los Angeles-Long Beach-Anaheim, CA (53% and 52.6%, respectively), and Miami-Fort Lauderdale-Pompano Beach, FL

metropolitan areas (52.9% and 50.9%, respectively). Similarly, both the proportion of Household Pulse Survey respondents that indicated that they received a COVID-19 vaccine and the proportion of tweets with positive sentiment were significantly higher in the Washington-Arlington-Alexandria Metro Area (55.3% and 58.1%, respectively) compared to the Dallas-Fort Worth-Arlington, TX (52.8% and 52.6%, respectively), Los Angeles-Long Beach-Anaheim, CA (53% and 52.6%, respectively), Miami-Fort Lauderdale-Pompano Beach, FL (52.9% and 50.9%, respectively), New York-Newark-Jersey City, NY-NJ-PA (52.5% and 55.3%, respectively), and Phoenix-Mesa-Chandler, AZ metropolitan areas (50.2% and 53.7%, respectively).

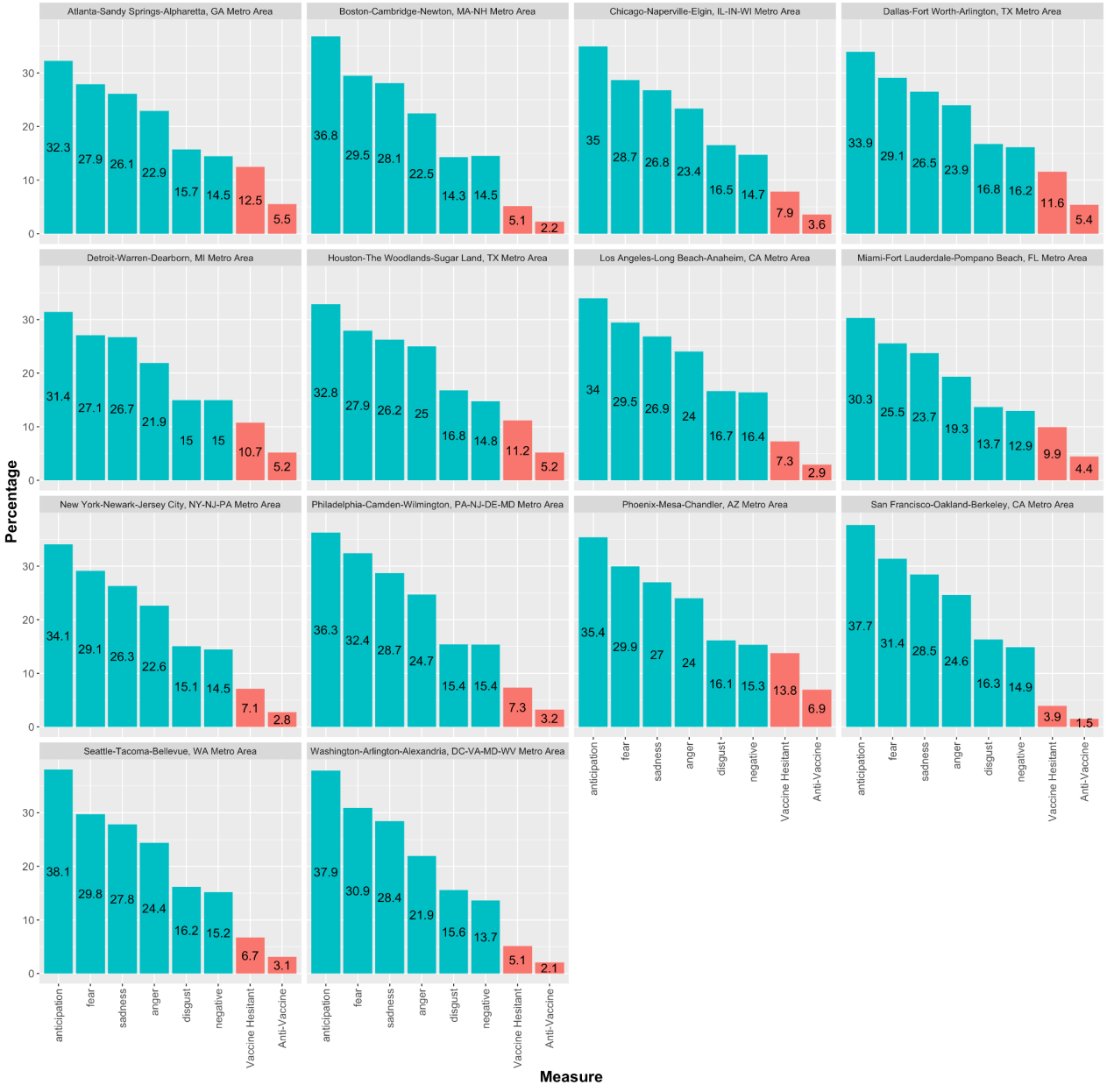
4.3.4.2 Vaccine Hesitant Measures

Figure 4.9 illustrates the proportion of negative sentiments & negative emotions found in tweets, compared to the proportion of Household Pulse Survey respondents that reported being anti-vaccine or vaccine hesitant. In this comparison, I used negative emotions & negative sentiments expressed in tweets as a proxy to measure vaccine hesitancy among Twitter users. Sentiments & emotions extracted from Tweets are shaded blue, while vaccine hesitant measures from Pulse Survey data are shaded red.

Across some metropolitan areas, the percentage of COVID-19 vaccine tweets expressing negative sentiment was closely aligned with the percentage of survey respondents that indicated they would “definitely NOT” or “probably NOT” get a vaccine once a vaccine to prevent COVID-19 was available to them. The Phoenix-Mesa-Chandler & Atlanta-Sandy Springs-Marietta metropolitan areas showed the closest alignment between negative COVID-19 vaccine perceptions across the two different data sources, with a difference of 1.5 and 2 percentage points, respectively, between the percentage of COVID-19 vaccine tweets expressing negative

sentiment (15.3% & 14.5%, respectively) and the percentage of survey respondents that indicated they would “definitely NOT” or “probably NOT” get a vaccine once a vaccine to prevent COVID-19 was available to them (13.8% & 12.5%, respectively). Other negative emotions, such as anticipation, fear, sadness, anger, and disgust, did not appear to align with any measures of COVID-19 vaccine hesitancy in the Household Pulse Survey data.

Figure 4.9 Comparison of vaccine hesitancy in Twitter data versus HPS data. Sentiments & emotions extracted from tweets are shaded blue, while measures from HPS data are shaded red.



The results of the proportions tests revealed very little alignment between the significant differences of the Twitter measure of vaccine hesitancy (negative sentiment) and the percentage of Household Pulse Survey respondents that were vaccine-hesitant (i.e. indicated that they *probably* or *definitely* would not get a COVID-19 vaccine once it became available to them). Both the proportion of Household Pulse Survey respondents that were vaccine hesitant and the proportion of tweets with negative sentiment were significantly lower in the Washington-Arlington-Alexandria Metro Area (5.1% and 13.7%, respectively) compared to the Dallas-Fort Worth-Arlington, TX Metro Area (11.6% and 16.1%, respectively) and Los Angeles-Long Beach-Anaheim, CA Metro Area (7.3% and 16.4%, respectively).

4.3.5. Predicting HPS vaccine attitudes using Twitter-based attitudes

The Model 1 regression analysis revealed significant associations between the percentage of vaccine compliant HPS respondents and the percentage of tweets expressing positive sentiment, joy, trust, and anticipation [Table 4.8]. The value of the coefficient of determination (R-squared) for the vaccine compliant model (Model 1) was 61.17%. This means approximately 61% of the variability in the percentage of vaccine compliant HPS respondents (dependent variable) was explained by the independent variables (percentage of tweets expressing positive sentiment, joy, surprise, trust, and anticipation) in the multiple linear regression model, which suggests we can predict fairly well vaccine compliance in the HPS using positive sentiments and emotions found on Twitter. The results show that an increase in the percentage of tweets expressing positive sentiment or anticipation (both $P<0.05$) is associated with an increase in the percentage of vaccine compliant HPS respondents. Contrastingly, an

increase in the percentage of tweets expressing joy or trust (*P*=.01 and *P*<0.05, respectively) is associated with a decrease in the percentage of vaccine compliant HPS respondents.

The Model 2 regression analysis revealed significant associations between the percentage of vaccine hesitant HPS respondents and the percentage of tweets expressing negative sentiment, anger, and anticipation. The R-squared value for the vaccine hesitant model (Model 2) was higher than the vaccine compliant model, at 72.16%. The independent variables (percentage of tweets expressing negative sentiment, anger, disgust, sadness, fear, and anticipation) in the multiple linear regression model explained approximately 72% of the variability in the percentage of HPS respondents that were vaccine hesitant, which suggests we can predict fairly well vaccine compliance in the HPS using negative sentiments and emotions found on Twitter. The results show that an increase in the percentage of tweets expressing anticipation (*P*<0.05) is associated with an increase in the percentage of vaccine hesitant HPS respondents. Contrastingly, an increase in the percentage of tweets expressing negative sentiment or anger (both *P*<0.05) is associated with a decrease in the percentage of vaccine hesitant HPS respondents.

Table 4.8 Linear Regression Model Results. Statically significant results (alpha = 0.05) are marked by an asterisk (*).

Model	Features	β coefficient	P-Value	Standard Error	R-squared
Model 1	% Positive	3.4221	5.46e-15 *	0.3759	61.17%
	% Joy	-1.8232	0.011792 *	0.7116	
	% Surprise	-0.5250	0.362716	0.5743	
	% Trust	-1.3539	0.000357 *	0.3671	
	% Anticipation	2.8030	9.87e-07 *	0.5396	

Model 2	% Negative	-0.36785	9.71e-05 *	0.09079	72.16%
	% Anger	-0.33164	0.000407 *	0.09083	
	% Disgust	0.09839	0.332028	0.10096	
	% Sadness	0.16890	0.099782	0.10172	
	% Fear	0.08012	0.400713	0.09496	
	% Anticipation	0.26032	1.02e-05 *	0.05616	

4.4. Discussion

4.4.1. Principal Findings

In this study, I sought out to validate social media as a data source by comparing the sentiments and emotions found in COVID-19 vaccine tweets to those expressed in the Census Bureau's Household Pulse Survey. A comparison of the public perceptions of COVID-19 vaccines found in tweets to those reflected in the Household Pulse Survey across 14 metropolitan areas in the United States revealed similar proportions of vaccine compliant sentiments in tweets and among survey respondents. Additionally, pairwise proportion tests for differences in vaccine compliant measures in tweets and the Household Pulse Survey revealed similar statistically significant differences. This study also examined whether the sentiments and emotions found in COVID-19 vaccine tweets can predict the vaccine hesitancy and compliance expressed in the Census Bureau's Household Pulse Survey. A linear regression analysis showed significant relationships between (1) the percentage of vaccine compliant HPS respondents and the percentage of tweets expressing positive sentiment, joy, trust, and anticipation; and (2) the percentage of vaccine hesitant HPS respondents and the percentage of tweets expressing negative sentiment, anger, and anticipation.

4.4.2. Study Findings in Context

The value of the coefficient of determination (R-squared) for both the vaccine compliant model (Model 1 – 61.17%) and vaccine hesitant model (Model 2 – 72.16%) suggests that vaccine perceptions in the HPS can be predicted fairly well using sentiments and emotions found on Twitter. The main objective of this study was to examine if aggregate attitudes extracted from social media can predict vaccine attitudes collected via surveys – more specifically, I hypothesized that there is a direct, positive relationship between (1) positive sentiments found in Twitter data and the HPS survey, and (2) negative sentiments found in Twitter data and the HPS survey.

The results of the linear regression revealed – as hypothesized – significant relationships between (1) the percentage of pro-vaccine HPS respondents and the percentage of tweets expressing positive sentiment, joy, trust, and anticipation; and (2) the percentage of anti-vaccine HPS respondents and the percentage of tweets expressing negative sentiment, anger, and anticipation. However, the direction of some of the relationships revealed in the linear regression models is not what I would expect. For example, I would expect to see a positive relationship between the positive sentiments and emotions on Twitter and vaccine compliance in the HPS, as suggested in a previous study that showed a positive relationship between positive sentiment scores in COVID-19 vaccine-related tweets and an increase in vaccination rates [102]. However, the regression model revealed a significant inverse relationship between the vaccine compliant measure in the HPS and the percentage of tweets expressing joy or trust. These findings might be indicative of individuals who are vaccine compliant and express these positive sentiments on Twitter, but have not received the vaccine just yet, for various reasons.

For example, data collection for this study started in early January 2021, but vaccine eligibility for adults in the United States was not expanded until mid-April 2021. Therefore, many individuals who were tweeting about the vaccine in a positive way were not able to get the vaccine during most of the study period due to eligibility reasons [103].

I would also expect to see a positive relationship between the negative sentiments and emotions on Twitter and vaccine hesitancy in the HPS. However, the regression model revealed a significant inverse relationship between the vaccine hesitant measure in the HPS and the percentage of tweets expressing negative sentiment or anger. These findings might be indicative of individuals whose online personas do not match their reality. For example, an individual might be obligated to get a vaccine due to their job or upcoming travel, making them vaccine compliant – but rant about it online. In our sample data, this type of person would be classified as "pro-vaccine" instead of "anti-vaccine" in the HPS but would also contribute to the negative perceptions found on Twitter. These findings also align with prior research that suggests an individual's online persona may differ from their offline identity [104]–[106]. This offline identity is oftentimes limited by physical, emotional, and financial circumstances that may be beyond an individual's control [105], [107]–[109]. However, individuals have complete control over the identity they choose to present online [104]–[106]. The inverse relationship between the vaccine hesitant measure in the HPS and the percentage of tweets expressing negative sentiment or anger may also be due to the use of sarcasm in tweets, where the text itself contradicts what is actually meant by the user [110].

The findings of this study contribute to the literature in two different ways. First, although many studies have examined COVID-19 vaccine acceptance by extracting information

from either surveys or social media, to my knowledge, there are no studies that provide a comparison and evaluate the relationship between these vastly different data sources. Unlike social media data, surveys come with postage, paper, printing, and data entry costs, making them costly to administer [19]. Evaluating the relationship between the attitudes found in surveys and on social media allows researchers to determine whether social media data can be trusted to reveal the same information we can extract from traditional surveys, or if I run the risk of losing important information just to cut costs. In this study, I found that COVID-19 vaccine attitudes in the HPS, measured as vaccine compliance and vaccine hesitancy, can be predicted using social media attitudes towards vaccines, measured via sentiments and emotions towards vaccines. The results of this study support the efforts of researchers, who over the past few years have looked to social media as a data source instead of traditional surveys, citing the availability of more readily available data and no or low cost data collection efforts [18].

The present study makes further contributions by revealing the sentiments and emotions found in tweets across different metropolitan areas. This builds upon several other studies that leveraged NLP methods such as sentiment analysis, emotion analysis, and topic modeling in order to examine vaccine-related perceptions [32]–[34]. In this study, I found most tweets expressed pro-vaccine sentiment, across all metropolitan areas. However, many tweets also expressed negative feelings and anticipation. This supports previous work, where researchers found lots of discussion about vaccine hesitancy, but ultimately found most tweets to have positive sentiment [35]. The present study also revealed trust as the dominant emotion found in tweets. This supports the results of a prior study that also found trust to be the

dominant emotion expressed in tweets, during an earlier time period [37]. A comparison of these results shows the vaccine conversation on Twitter remained relatively consistent over time.

The present study provides further evidence of the benefits of using social media data for public health research. The overarching contribution of this work suggests the adaptation of alternative data sources and natural language processing techniques to assist in public health decision making.

4.4.3. Limitations and Future Work

Considering the limitations of the present study may lead to future, related work. This study places an emphasis on using Twitter as a data source, but the lack of representation among Twitter users leads to bias in the sample. For example, Twitter users tend to be younger, more educated, have higher incomes, and more liberal [16]. The lack of representation among Twitter users suggests limited generalizability of the results to the larger population. Adding to this lack of representation is the limited sample of tweets available to the public via the Twitter Streaming API, which makes available a random sample of 1% of all Tweets sent by Twitter users at any given time [14]. Additionally, in a study that sought to assess the perceptions of the COVID-19 vaccine, individuals who do not have access to social media are systematically excluded from the analysis sample. Requiring tweets to have some type of location information further limits the tweets sample and poses cause for concerns of bias in results. For example, there may be concerns about systematic differences in tweets with and without location information. However, the comparative analysis of sentiments and emotions in tweets with and without location information showed close alignment of vaccine perceptions.

Future studies should endeavor to utilize other NLP approaches, such as topic modeling, to compare the public perceptions of the COVID-19 vaccine on social media to those found in surveys. The survey used in this study, the Household Pulse Survey, presented respondents with in-depth questions related to why they were vaccine hesitant, so applying topic models to tweets may reveal some of the same attitudes and themes as those expressed in the survey. Future studies may also involve pulling data from other social media platforms, such as Facebook, and comparing the overall perceptions reflected across all mediums.

4.4.4. Conclusions

The ongoing COVID-19 pandemic requires consistent monitoring and data driven public health policies. To slow the spread of the virus, public health officials have stressed that vaccines are essential in the world-wide battle against COVID-19. However, vaccine hesitancy continues to be a barrier for effective and consistent vaccine rollout programs. Prior efforts have utilized surveys to gauge attitudes towards the COVID-19 vaccine, but this study suggests that these public perceptions may also be extracted from a readily available, low-cost data source – social media. In this study, I validated social media as a data source by evaluating the relationship between the attitudes expressed among Twitter users and respondents to the Household Pulse Survey as well as the ability for attitudes expressed among Twitter users to predict vaccine compliance and hesitancy among HPS respondents.

Chapter 5 – Study #3: Using COVID-19 vaccine Twitter chatter to predict vaccination rates in the United States

5.1. Introduction

5.1.1. Background

Since the onset of the COVID-19 pandemic, there has been a global effort to develop vaccines that protect against COVID-19. Individuals who are fully vaccinated are far less likely to contract and therefore transmit the virus to others [111]. Up until recently, public health experts have stressed the importance of achieving a numerical threshold of herd immunity, but this is only possible if a significant proportion of the population is fully vaccinated. More recent research suggests that the traditional concept of herd immunity may not apply to COVID-19 [112]. Instead, the goal is to increase vaccination uptake to optimize population protection without prohibitive restrictions on our daily lives [113]. Accurately forecasting vaccination uptake allows policy makers and researchers to evaluate how close we are to achieving normalcy again.

Researchers have turned to traditional methods for forecasting COVID-19 infection and vaccination rates [114]–[116]. For example, one of the most common forecasting methods used, univariate time series, involves predicting future vaccination rates using historical vaccination rates. While this method can be useful in many cases, it fails to account for other time-dependent factors that may also influence vaccinations. For example, The COVID-19 vaccine conversation on social media has been deemed an infodemic, with anti-vaccination misinformation spreading across social media platforms [117]. Researchers have found that the

internet and social media both play a role in shaping personal or parental choices about vaccinations [118], [119]. Additionally, previous research showed a positive relationship between positive sentiment scores in COVID-19 vaccine-related tweets and an increase in vaccination rates [102]. These finding suggest it is important to consider the daily conversations on social media when developing vaccine uptake forecast models.

5.1.2. Forecasting COVID-19 Related Measures Using Social Media

There is no shortage of studies that sought to forecast COVID-19-related measures using information from social media. Researchers in [39] conducted a study using COVID-19 related terms mentioned in tweets and Google searches to predict COVID-19 waves in the United States. Researchers found that tweets that mentioned COVID-19 symptoms predicted 100% of first waves of COVID-19 days sooner than other data sources. Another study used data from Google searches, tweets, and Wikipedia page views to predict COVID-19 cases and deaths in the United States [40]. Researchers found models that included features from all three sources performed better than baseline models that did not include these features. Researchers also found that Google searches were a leading indicator of the number of cases and deaths across the United States. Another study [41] examined the relationship between daily COVID-19 cases and COVID-19 related tweets and Google Trends. In a study conducted by [42], researchers used reports of symptoms and diagnoses on Weibo, a popular social media platform in China, in order to predict COVID-19 case counts in Mainland China. Researchers found reports of symptoms and diagnoses on the social media platform to be highly predictive of daily case counts. Although each of these studies forecast COVID-19 cases and deaths, none of these studies forecast COVID-19 vaccination rates.

5.1.3. Forecasting Vaccinations

Very few studies have conducted time series forecasting of the COVID-19 vaccinated population in the United States. In a study conducted by [32], researchers developed a time series model to predict the percentage of the US population that would get at least one dose of the COVID-19 vaccine or be fully vaccinated. Researchers projected that by the end of July 2021, 62.44% and 48% of the US population would get at least one dose of the COVID-19 vaccine or be fully vaccinated, respectively. Although this paper also included a separate tweet sentiment analysis, researchers did not include Twitter-related features in the forecast model. Additionally, researchers used aggregated vaccination data for the entire United States, rather than a more granular geographic level.

Another study aimed to evaluate if and when the world would reach a vaccination rate sufficient enough for herd immunity by forecasting the number of people fully vaccinated against COVID-19 in various countries, including the US [43]. In this study, researchers used a common univariate time series forecasting method, Autoregressive Integrated Moving Average (ARIMA), to forecast the future number of fully vaccinated people using only historical vaccination data. Based on the resulting projections, researchers concluded that countries were nowhere near the necessary herd immunity threshold needed to end the COVID-19 pandemic.

A study conducted by [44] sought to predict COVID-19 vaccine uptake using various sociodemographic factors. Although not a time series forecasting model, the results of this study showed that geographic location, education level, and online access were highly predictive of vaccination uptake in the United States. The model predicted vaccine uptake with 62% accuracy.

Although there are very few studies related to COVID-19 vaccination forecasting, other studies have been conducted to predict immunizations for other illnesses. For example, one study analyzed electronic medical records of a cohort of 250,000 individuals over the course of ten years [45]. Researchers developed a model to predict vaccination uptake of individuals in the upcoming influenza season based on previous personal and social behavioral patterns. Another study developed a tool for leveraging immunization related content from Twitter and Google Trends to develop a model for predicting whether a child would receive immunizations [46]. Researchers were able to predict child immunization status with 76% accuracy.

5.1.4. Study Objectives

Although few previous studies have developed forecast models for COVID-19 vaccination rates in the United States, to our knowledge, there are no studies that aim to factor in the real-time vaccination attitudes present on Twitter. The results of Study #2 (Chapter 4) showed that attitudes toward COVID-19 vaccines found in tweets were predictive of vaccine attitudes in the Household Pulse Survey. These findings led to the motivation for the present study, where I hypothesize that using vaccine attitudes on Twitter as features in vaccine uptake forecast models may improve the performance of these forecast models. Previous studies developed forecast models that focused on the entire United States as a whole. These forecast models fail to appreciate the differences in vaccination roll out, behaviors, and attitudes across different geographic regions. The present study also seeks to fill this gap by examining vaccine uptake at the metropolitan level.

The purpose of this study is to develop a time series forecasting algorithm that can predict future vaccination rates across US metropolitan areas. Specifically, the present study

aims to determine whether supplementing forecast models with real-time vaccine attitudes found in tweets – measured via sentiments and emotions – improves over baseline models that only use historical vaccination data. Developing a predictive tool for vaccination uptake in the United States will empower public health researchers and decision makers to design targeted vaccination campaigns in hopes of achieving the vaccination threshold required to reach herd immunity.

5.2. Materials and Methods

5.2.1. Data Collection and Preprocessing

5.2.1.1. Twitter Data

The Twitter Streaming Application Programming Interface (API), which provides access to a random sample of 1% of publicly available tweets, was used to collect tweets from 8 of the top 10 most populated metropolitan areas in the United States from January 2021 to May 2021 [Table 5.1] [120]. I chose to focus on large metropolitan areas in order to gather a sufficient number of tweets for the analysis. Additionally, larger metropolitan areas also tend to have users who enable the location feature when tweeting [121], [122]. It is important to note that although the Dallas-Fort Worth-Arlington, TX & Houston-The Woodlands-Sugar Land, TX metropolitan areas are among the 10 most populated metropolitan areas in the United States, tweets from these areas were not included in the analysis due to the lack of available vaccination data during the study period – data that was required for the forecast models.

All tweets had “place” information (usually city and state). The place information found in tweets was used to determine the metropolitan area associated with each tweet. Next, to extract tweets related to COVID-19 vaccines, tweets were further filtered by matching

variations of vaccine-related keywords, such as *vaccine, pfizer, moderna, johnson & johnson, and dose*. A complete list of vaccine-related keywords can be found in Appendix B. The tweets sample was further preprocessed to minimize "noise" resulting from tweets that matched our vaccine-related keywords but did not necessarily reflect the thoughts and opinions of individual Twitter users. For example, companies often promote job postings and advertisements on Twitter using targeted hashtags in hopes of reaching their target audience. To prevent these tweets from adding noise to the sample, tweets related to job postings and advertisements were removed by excluding tweets with hashtags and keywords such as "#jobs", "#hiring", and "#ad".

Table 5.1 Targeted metropolitan areas for Twitter data collection, January 1 - May 20, 2021.

Phoenix-Mesa-Chandler, AZ	Philadelphia-Camden-Wilmington, PA-NJ-DE-MD
Miami-Fort Lauderdale-Pompano Beach, FL	Washington-Arlington-Alexandria, DC-VA-MD-WV
Atlanta-Sandy Springs-Alpharetta, GA	Chicago-Naperville-Elgin, IL-IN-WI
New York-Newark-Jersey City, NY-NJ-PA	Los Angeles-Long Beach-Anaheim, CA

5.2.1.2. COVID-19 Vaccination Data

Daily COVID-19 vaccination data at the county-level was collected for the January 2021 to May 2021 study period from the Centers for Disease Control and Prevention's (CDC) publicly available vaccination dataset [123]. This dataset includes daily vaccination data from clinics, pharmacies, long-term care facilities, dialysis centers, Federal Emergency Management Agency and Health Resources and Services Administration partner sites, and federal entity facilities. Vaccination administration data are reported to the CDC via immunization information systems (IISs), the vaccine administration management system (VAMS), and data submissions directly to the COVID-19 Data Clearing House [123]. Each county was linked to its respective metropolitan area according to the US Census delineation file [124]. Next, the data was aggregated to the

daily-metropolitan level and the seven-day rolling average of the percentage of individuals who have been administered at least one vaccine dose was calculated.

5.2.2. Data Analysis

5.2.2.1. Sentiment and Emotion Analysis of Tweets

For the purposes of this study, I measure COVID-19 vaccine attitudes via sentiment and emotion analyses of tweets. To capture the sentiments and emotions found in COVID-19 vaccine-related tweets, a sentiment and emotion analysis of all tweets was conducted using the NRC lexicon from the Syuzhet package in R [74]. The NRC lexicon, developed by Saif Mohammad, contains a list of manually labeled English words and their associations with negative and positive sentiments and common human emotions, such as trust, fear, sadness, surprise, and disgust [98]. The Syuzhet package applies the NRC lexicon by independently evaluating and rating each word or expression within a tweet [99]. The get_nrc_sentiment function was applied to all tweets to calculate the valence of eight different emotions (fear, joy, anticipation, anger, disgust, sadness, surprise, trust), along with overall positive and negative sentiment. As mentioned in Chapter 4, the accuracy of the sentiment classifier, based on a random sample of 1000 tweets, was 73.4%.

The percentages of the eight emotions, along with the percentage of positive, neutral, and negative sentiments were calculated at the metropolitan level. The total number of COVID-19 vaccine related tweets and users per 100,000 population was also calculated for each day of data collection, at the metropolitan level. Finally, user engagement metrics, including the average number of re-tweets and favorites, were calculated for each day of data collection, at the metropolitan level. Retweets and favorites suggest, after processing the information, that a

user resonates with an idea expressed in a tweet [125], [126]. Therefore, I believe these engagement metrics might also reflect vaccine attitudes.

5.2.2.2. Time Series Model

The data were divided into training and test datasets, where the time series analysis was trained using the dataset created from the January 1 - April 12, 2021 time period, and tested on the dataset created from the April 13th - May 20th, 2021 time period. Auto-Regressive Integrated Moving Average (ARIMA) models were executed for forecasting the proportion of individuals who have been administered at least one vaccine dose. ARIMAX models, which are extensions of ARIMA models that include independent predictors called exogenous variables, were also executed. The ARIMA method has been widely used in time series forecasting and public health surveillance [127]–[129]. An ARIMA model typically consists of three components: (1) auto-regression, notated in the model as p; (2) differencing, notated in the model as d; and (3) moving average, notated in the model as q [130]. In an ARIMA model, the present value of the time-series is a linear function of random noise and its previous values; the present value is also a linear function of both present and past values of the residuals in the model; and the auto-regressive moving average model includes both the auto-regressive and moving average models, in addition to the historical values in the time series and its residuals [127].

Stationarity of a time series is a key assumption when making predictions based on past observations of a variable [131]. Stationarity requires the properties (mean and variance) of a time series to remain constant over time, thus making future values easier to predict [132]. Otherwise, the results are spurious and analyses are not valid [126]. The stationarity of all variables included in the time series was assessed using the Dickey-Fuller (dfuller) test. If the

null hypothesis is rejected, stationarity is satisfied. If stationarity is not satisfied, variables must undergo differencing, a process that removes any trend in the times series that is not of interest [131]. All differencing and model selection was performed by the auto.arima function from the *forecast* package in R, which is a function that selects the optimal order of the model based on the Hyndman-Khandakar algorithm for automatic ARIMA modeling [132]. A combination of unit root tests and minimization of the AIC and MLE allows this algorithm to select the best preforming model order by fitting several variations of model components *p, d,* and *q [133]*.

For each metropolitan area, a baseline ARIMA model with no exogenous variables was constructed to forecast the seven-day rolling average of the number of individuals who have been administered at least one vaccine dose, using only past values of this outcome. To assess the ability of vaccine attitudes on Twitter to improve COVID-19 vaccination forecasts, multiple ARIMAX models were executed, each with individual Twitter-derived features included as exogenous variables. Additionally, I executed a multivariate ARIMAX model that included those Twitter attitudes that showed improvement over the ARIMA baseline across all metro areas. A final ARIMAX model that contained all Twitter features regardless of performance was attempted but did not converge. A complete list of the constructed time series models can be found in Table 5.2.

Table 5.2 Time series models predicting COVID-19 vaccine uptake, January 1 – May 20, 2021.

Model Type	Exogenous Variable(s)
ARIMA	None (baseline)
ARIMAX	Number of users per 100,000 population
ARIMAX	Number of tweets per 100,000 population
ARIMAX	Average favorites
ARIMAX	Average retweets
ARIMAX	% Positive Sentiment
ARIMAX	% Negative Sentiment
ARIMAX	% Neutral Sentiment
ARIMAX	% Trust
ARIMAX	% Surprise
ARIMAX	% Sadness
ARIMAX	% Joy
ARIMAX	% Fear
ARIMAX	% Disgust
ARIMAX	% Anticipation
ARIMAX	% Anger
ARIMAX	Best Predictors (predictors that show improvement over baseline across all metro areas)

5.3. Results

5.3.1. Twitter Data

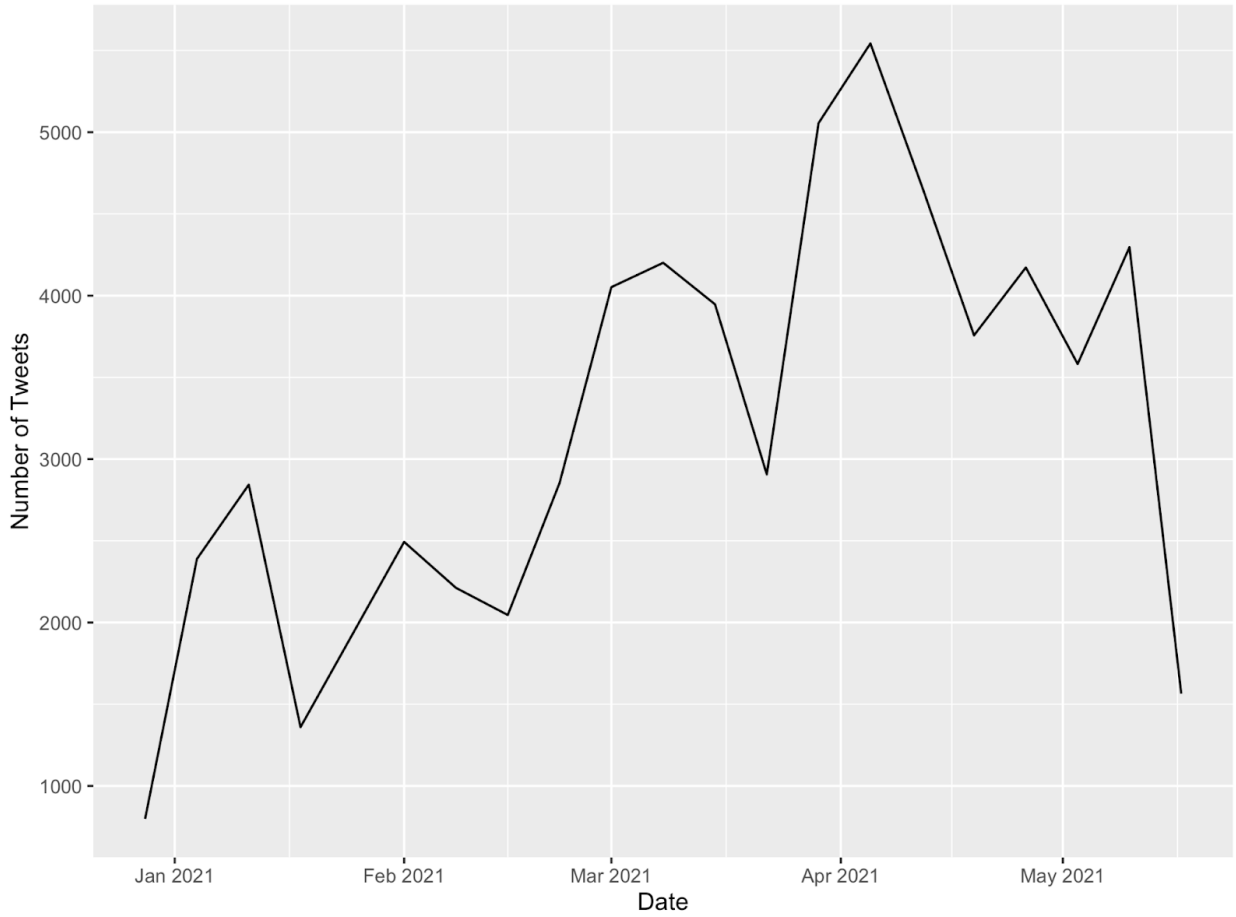
A total of 64,737 COVID-19 vaccine-related tweets were collected during the data collection period, across 25,905 users [Table 5.3]. The Los Angeles-Long Beach-Anaheim metropolitan area had the largest representation of tweets (13,125/64,737, 20.27%), as well as the largest representation of users (5,620/25,905, 21.69%). The Houston-Woodlands-Sugar Land metropolitan area had the smallest representation of tweets (999/64,737, 1.54%) as well as the smallest representation of users (541/25,905, 2.09%). The maximum number of tweets by a single individual was 228 (from a user in the Washington-Arlington-Alexandria metropolitan area).

The temporal trends for the number of COVID-19 vaccine-related tweets from January to May 2021 are presented in Figure 5.1. The number of COVID-19 vaccine-related tweets fluctuated over time; however, a peak in the number of tweets was observed during the week of April 5-11th, 2021. This was the week that President Joe Biden announced that every adult in the United States would be eligible to receive a COVID-19 vaccine starting April 19, 2021 [102].

Table 5.3 Number of COVID-19 vaccine tweets (N=64,737) and users (N=25,905) by city, January 1 – May 20, 2021.

Metropolitan Area	Number of tweets, n (%)	Number of users, n (%)	Average Retweets, mean (sd)	Average Favorites, mean (sd)
Atlanta-Sandy Springs-Alpharetta, GA	12623 (19.5)	5431 (20.97)	438 (5140)	10 (178)
Chicago-Naperville-Elgin, IL-IN-WI	6857 (10.59)	2847 (10.99)	543 (9579)	11 (118)
Los Angeles-Long Beach-Anaheim, CA	13125 (20.27)	5620 (21.69)	438 (6415)	16 (174)
Miami-Fort Lauderdale-Pompano Beach, FL	1631 (2.52)	625 (2.41)	176 (1891)	10 (91)
New York-Newark-Jersey City, NY-NJ-PA	12387 (19.13)	4858 (18.75)	351 (4209)	13 (224)
Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	4345 (6.71)	1558 (6.01)	267 (3187)	131 (2389)
Phoenix-Mesa-Chandler, AZ	2231 (3.45)	914 (3.53)	169 (1704)	6 (20)
Washington-Arlington-Alexandria, DC-VA-MD-WV	6488 (10.02)	2025 (7.82)	304 (3952)	13 (124)

Figure 5.1 Number of COVID-19 vaccine tweets over time, across all metropolitan areas, January 1 – May 20, 2021



5.3.2. Sentiment & Emotion Analysis

A sentiment analysis classified most tweets across all metropolitan areas as having positive sentiment, with trust as the predominantly expressed emotion [Table 5.4]. The Washington-Arlington-Alexandria metropolitan area had the largest proportion of tweets with positive sentiment (57.8%), while the Miami-Ft. Lauderdale-Pompano Beach metropolitan area had the lowest proportion of tweets with positive sentiment (50.5%). Tweets with negative sentiment held the smallest proportions across all metropolitan areas, with sadness and fear being the most perceived negative emotions. The Phoenix-Mesa-Chandler metropolitan area had the largest proportion of tweets with negative sentiment (17.2%), while the Miami-Ft.

Lauderdale-Pompano Beach metropolitan area had the lowest proportion of tweets with negative sentiment (12.8%). Tweets with neutral sentiment represented between ~29% and 37% of tweets across all metropolitan areas.

Table 5.4 Distribution of sentiments and emotions among COVID-19 vaccine tweets collected from January 1 – May 20, 2021 (N=64,737).

Metropolitan Area	% Positive	% Negative	% Neutral	% Trust	% Anticipation	% Sadness	% Anger	% Fear	% Joy	% Surprise	% Disgust
Atlanta-Sandy Springs-Alpharetta, GA	53.7	15.5	30.9	43.0	33.1	26.2	24.0	28.6	24.0	21.2	16.1
Chicago-Naperville-Elgin, IL-IN-WI	55.4	15.3	29.4	44.6	34.3	26.3	23.6	27.9	26.3	22.6	16.2
Los Angeles-Long Beach-Anaheim, CA	51.8	16.4	31.7	42.3	33.9	26.8	24.2	29.5	26.2	22.0	16.4
Miami-Fort Lauderdale-Pompano Beach, FL	50.5	12.8	36.7	40.8	30.6	23.4	19.4	25.3	23.7	18.3	13.7
New York-Newark-Jersey City, NY-NJ-PA	55.6	14.1	30.3	43.7	34.0	25.9	23.0	29.2	26.6	21.7	15.3
Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	55.0	15.7	29.3	47.2	35.8	29.0	24.8	32.4	26.4	22.9	15.9
Phoenix-Mesa-Chandler, AZ	53.2	17.2	29.5	45.3	36.1	28.4	26.4	33.0	25.5	24.2	16.0
Washington-Arlington-Alexandria, DC-VA-MD-WV	57.8	13.5	28.7	48.1	37.3	28.3	22.3	31.2	28.6	23.5	15.8

5.3.2. Time Series Forecast

Multiple time series models were constructed to forecast the vaccine uptake rate (7-day rolling average). The results of the Dickey-Fuller (dfuller) test for stationarity revealed that in some metropolitan areas, stationarity did not hold for the outcome variable, vaccination rate, and several of the exogenous variables, including number of users per 100,000 population, number of tweets per 100,000 population, and percentage of tweets expressing trust, anticipation, anger, and fear [Table 5.5]. However, the necessary differencing was automatically applied via the auto.arima function.

Table 5.5 Dickey-Fuller (dfuller) Test for Stationarity. Non-stationary variable results are marked by an asterisk (*).

	Phoenix-Mesa-Chandler, AZ		**Miami-Fort Lauderdale-Pompano Beach, FL**		**Atlanta-Sandy Springs-Alpharetta, GA**		**Philadelphia-Camden-Wilmington, PA-NJ-DE-MD**	
Variable	**Test Statistic**	**p-value**	**Test Statistic**	**p-value**	**Test Statistic**	**p-value**	**Test Statistic**	**p-value**
% of individuals who have been administered at least one vaccine dose (7 day rolling average)	-0.93	0.95*	-1.34	0.85*	-2.21	0.49*	-1.58	0.75*
Number of users per 100,000 population	-2.72	0.28*	-3.15	0.1*	-2.1	0.54*	-2.21	0.49*
Number of tweets per 100,000 population	-3.21	0.09*	-3.13	0.11*	-2.43	0.4*	-2.72	0.28*
Average favorites	-4.97	0.01	-4.73	0.01	-8.75	0.01	-4.55	0.01
Average retweets	-4.4	0.01	-4.7	0.01	-4.53	0.01	-5.71	0.01
% Positive Sentiment	-3.75	0.02	-3.74	0.02	-3.67	0.03	-3.77	0.02
% Negative Sentiment	-4.21	0.01	-4.36	0.01	-3.56	0.04	-4.45	0.01
% Neutral Sentiment	-4.89	0.01	-5.05	0.01	-3.72	0.03	-4.75	0.01

% Trust	-2.99	0.17*	-3.04	0.14*	-3.11	0.12*	-2.98	0.17*
% Surprise	-4.96	0.01	-4.25	0.01	-4.41	0.01	-3.78	0.02
% Sadness	-4.6	0.01	-5.03	0.01	-4.8	0.01	-4.19	0.01
% Joy	-4.01	0.01	-3.86	0.02	-4.81	0.01	-4.04	0.01
% Fear	-5.63	0.01	-4.83	0.01	-4.61	0.01	-3.93	0.01
% Disgust	-5.6	0.01	-4.95	0.01	-4.16	0.01	-5.22	0.01
% Anticipation	-4.98	0.01	-3.01	0.16*	-3.71	0.03	-5.23	0.01
% Anger	-5.71	0.01	-4.63	0.01	-4.55	0.01	-4.52	0.01
	Washington-Arlington-Alexandria, DC-VA-MD-WV		**Chicago-Naperville-Elgin, IL-IN-WI**		**Los Angeles-Long Beach-Anaheim, CA**		**New York-Newark-Jersey City, NY-NJ-PA**	
Variable	**Test Statistic**	**p-value**	**Test Statistic**	**p-value**	**Test Statistic**	**p-value**	**Test Statistic**	**p-value**
% of individuals who have been administered at least one vaccine dose (7 day rolling average)	-1.64	0.72*	-1.69	0.71*	-1.91	0.62*	-1.55	0.76*
Number of users per 100,000 population	-2.84	0.23*	-2.49	0.37*	-1.36	0.84*	-2.44	0.4*
Number of tweets per 100,000 population	-2.76	0.26*	-2.88	0.21*	-1.68	0.71*	-2.33	0.44*
Average favorites	-4.4	0.01	-4.24	0.01	-4.19	0.01	-4.04	0.01
Average retweets	-5.52	0.01	-5.42	0.01	-5.2	0.01	-5.67	0.01
% Positive Sentiment	-5.13	0.01	-3.41	0.06*	-3.93	0.01	-4.3	0.01
% Negative Sentiment	-5.24	0.01	-3.95	0.01	-4.04	0.01	-4.32	0.01
% Neutral Sentiment	-4.73	0.01	-3.92	0.02	-4.35	0.01	-4.86	0.01
% Trust	-3.37	0.06*	-3.01	0.16*	-3.76	0.02	-4.17	0.01

% Surprise	-3.41	0.06*	-4.22	0.01	-4.11	0.01	-4.24	0.01
% Sadness	-3.55	0.04	-3.88	0.02	-4.57	0.01	-7.36	0.01
% Joy	-4.48	0.01	-4.35	0.01	-6.24	0.01	-3.73	0.02
% Fear	-4.74	0.01	-2.88	0.21*	-4.67	0.01	-4.1	0.01
% Disgust	-3.28	0.08*	-5.08	0.01	-6.09	0.01	-4.45	0.01
% Anticipation	-4.77	0.01	-3.93	0.02	-4.72	0.01	-4.63	0.01
% Anger	-3.72	0.03	-3.3	0.07*	-5.15	0.01	-5.49	0.01

The performance of the optimal models across all regions, as determined by the auto.arima function, can be found in Table 5.6. The best performing model for each metropolitan area is marked by an asterisk. Model performance for the "out-sample" forecasts was evaluated using the root mean square error (RMSE) instead of AIC because RMSE measures how close the data are around the line of best fit [135]. This measure is commonly used in time series forecasting to evaluate how close the forecasted values are to the actual values [135]. When evaluating model performance using RMSE, across all metropolitan areas, the addition of a Twitter-derived feature related to COVID-19 vaccination attitudes improved model performance by up to 9%. For example, in both the Phoenix-Mesa-Chandler & Atlanta-Sandy Springs-Alpharetta metropolitan areas, adding the *percentage of vaccine tweets expressing trust* as an exogenous variable not only resulted in a lower RMSE compared to the baseline ARIMA model, but also resulted in the lowest RMSE across all the models within these metropolitan areas. Additionally, in these two metropolitan areas, all the ARIMAX models, which each had one Twitter-derived feature related to COVID-19 vaccination attitudes, showed

improvement over the baseline ARIMA model that did not factor in Twitter-derived features. A final model that contained the 3 features that consistently showed improvement over baseline across all metro areas (% Negative Sentiment, % Surprise, % Sadness) showed improvement over the baseline ARIMA when combined into one model (ARIMAX with multiple exogenous variables). A final ARIMAX model that contained all Twitter features regardless of performance was attempted but did not converge.

Table 5.6 also shows the components (p, d, q) of the ARIMA models. A key component, the p component, represents the number of lag observations in the model – also known as the lag order. Across all metropolitan areas, most of the ARIMA/ARIMAX models had p=1, which means that vaccination rates and Twitter features from the previous day are best used to predict future vaccination rates. In the New York-Newark-Jersey City, NY-NJ-PA and Washington-Arlington-Alexandria, DC-VA-MD-WV metro areas, a few of the ARIMA/ARIMAX models had p=2, which means that vaccination rates and Twitter features (such as the average number of retweets and favorites) from two days prior are best used to predict future vaccination rates. In the Chicago-Naperville-Elgin, IL-IN-WI and Miami-Fort Lauderdale-Pompano Beach, FL metro areas, the ARIMAX models containing the average number of retweets had p=3, which means that vaccination rates and the average number of retweets from three days prior are best used to predict future vaccination rates.

Table 5.6 ARIMA/ARIMAX Model Performance (RMSE) and Components (p,d,q). Models that performed better than the baseline ARIMA are marked by an asterisk (*).

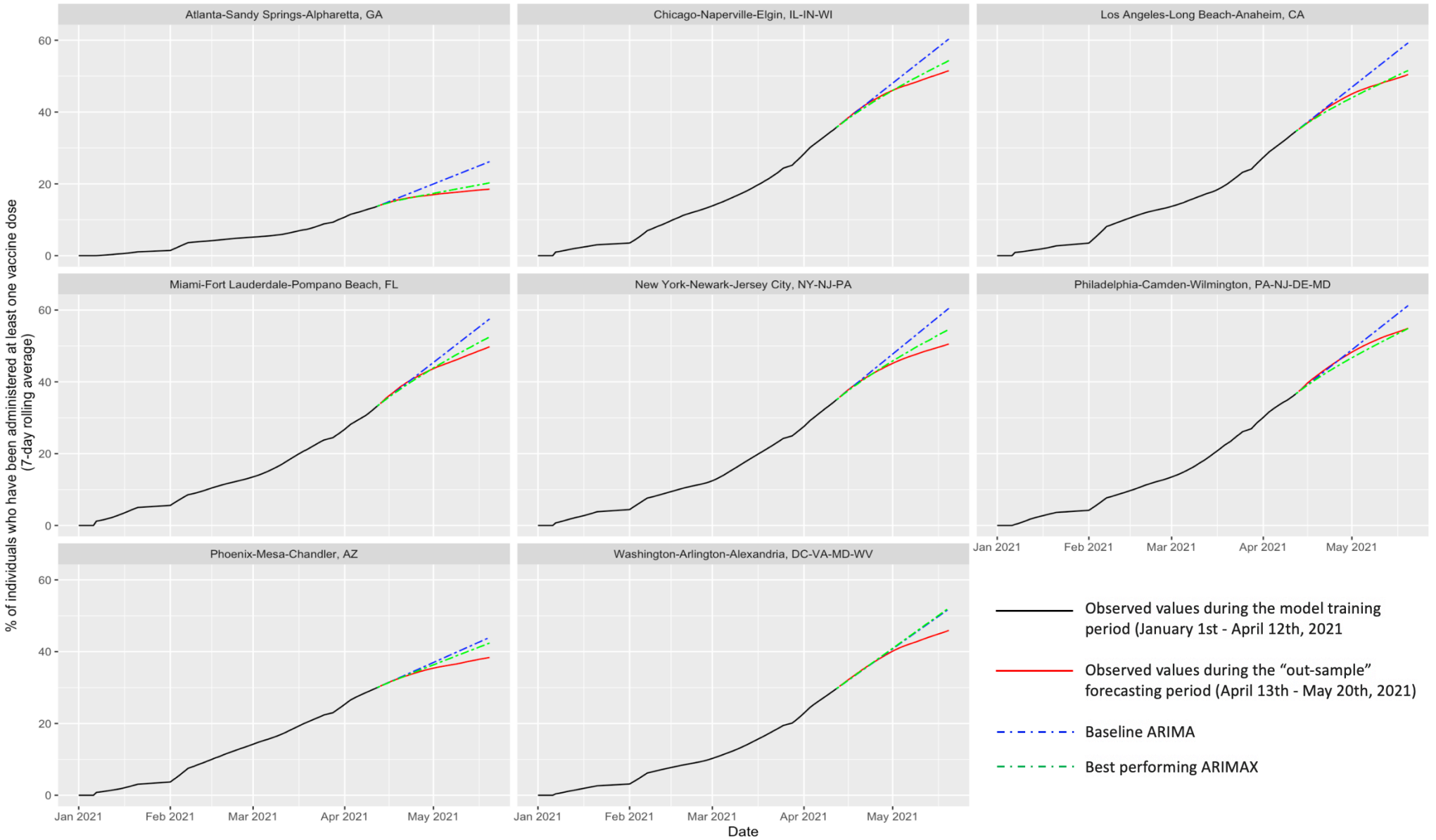
Variables	Phoenix-Mesa-Chandler, AZ		Miami-Fort Lauderdale-Pompano Beach, FL		Atlanta-Sandy Springs-Alpharetta, GA		New York-Newark-Jersey City, NY-NJ-PA	
	RMSE	p,d,q	RMSE	p,d,q	RMSE	p,d,q	RMSE	p,d,q
(Baseline) % of individuals who have been administered at least one vaccine dose (7 day rolling average)	0.1217	1, 1, 1	0.1516	1, 1, 1	0.0510	1, 1, 1	0.1039	1, 1, 1
Number of users per 100,000 population	0.1185	1, 1, 1	0.1466*	1, 1, 1	0.0498	1, 1, 0	0.1015	1, 1, 1
Number of tweets per 100,000 population	0.1188	1, 1, 1	0.1482	1, 1, 1	0.0497	1, 1, 0	0.1011*	1, 1, 1
Average favorites	0.1177	1, 1, 1	0.1516	0, 2, 1	0.0509	1, 1, 0	0.1055	2, 1, 0
Average retweets	0.1217	0, 2, 1	0.1567	3, 1, 0	0.0507	0, 2, 0	0.1042	2, 1, 0
% Positive Sentiment	0.1188	1, 1, 1	0.1518	1, 1, 1	0.0501	1, 1, 0	0.1034	1, 1, 1
% Negative Sentiment	0.1175	1, 1, 1	0.1516	0, 2, 1	0.0501	1, 1, 0	0.1036	0, 2, 1
% Neutral Sentiment	0.1191	1, 1, 1	0.1518	1, 1, 1	0.0502	1, 1, 0	0.1026	1, 1, 1
% Trust	0.1172*	1, 1, 1	0.1516	1, 1, 1	0.0464*	1, 1, 4	0.1035	1, 1, 1
% Surprise	0.1193	1, 1, 1	0.1511	0, 2, 1	0.0501	1, 1, 0	0.1037	1, 1, 1
% Sadness	0.1188	1, 1, 1	0.1503	1, 1, 1	0.0493	1, 1, 0	0.1038	1, 1, 1
% Joy	0.1192	1, 1, 1	0.1515	1, 1, 1	0.0469	1, 1, 4	0.1034	1, 1, 1
% Fear	0.1193	1, 1, 1	0.1512	0, 2, 1	0.0499	1, 1, 0	0.1036	1, 1, 1
% Disgust	0.1193	1, 1, 1	0.1518	1, 1, 1	0.0499	1, 1, 0	0.1038	1, 1, 1
% Anticipation	0.118	1, 1, 1	0.1484	1, 1, 1	0.0501	1, 1, 0	0.1028	1, 1, 1
% Anger	0.1193	1, 1, 1	0.1467	1, 1, 1	0.0504	0, 2, 0	0.1037	1, 1, 1
Best Predictors (% Negative Sentiment, % Surprise, % Sadness)	0.1175	1, 1, 1	0.1495	1, 1, 1	0.0499	1, 1, 0	0.1033	1, 1, 1

Variables	Philadelphia-Camden-Wilmington, PA-NJ-DE-MD		Washington-Arlington-Alexandria, DC-VA-MD-WV		Chicago-Naperville-Elgin, IL-IN-WI		Los Angeles-Long Beach-Anaheim, CA	
	RMSE	p,d,q	RMSE	p,d,q	RMSE	p,d,q	RMSE	p,d,q
(Baseline) % of individuals who have been administered at least one vaccine dose (7 day rolling average)	0.0856	1, 1, 1	0.0757	1, 1, 1	0.1365	1, 1, 1	0.1457	1, 1, 1
Number of users per 100,000 population	0.0848	1, 1, 0	0.0760	1, 1, 0	0.1351	1, 1, 1	0.1436	1, 1, 1
Number of tweets per 100,000 population	0.0848	1, 1, 0	0.0759	1, 1, 0	0.1355	1, 1, 1	0.144	1, 1, 1
Average favorites	0.0856	0, 2, 0	0.0759	2, 1, 0	0.1365	0, 2, 1	0.1458	1, 1, 1
Average retweets	0.0857	1, 1, 0	0.0754	2, 1, 0	0.136	3, 1, 0	0.1454	0, 2, 1
% Positive Sentiment	0.0844*	1, 1, 0	0.0752	2, 1, 0	0.1331	1, 1, 1	0.1433	1, 1, 1
% Negative Sentiment	0.0846	1, 1, 0	0.0752	1, 2, 0	0.1354	1, 1, 1	0.1441	1, 1, 1
% Neutral Sentiment	0.0847	0, 2, 0	0.0752	1, 1, 0	0.1304*	1, 1, 1	0.1433	1, 1, 1
% Trust	0.0846	1, 1, 0	0.0762	1, 1, 0	0.1361	1, 1, 1	0.1433	1, 1, 1
% Surprise	0.0856	0, 2, 0	0.0754	2, 1, 0	0.1359	1, 1, 1	0.1441	1, 1, 1
% Sadness	0.0853	0, 2, 0	0.0756	0, 2, 1	0.1359	1, 1, 1	0.1423	1, 1, 1
% Joy	0.0856	1, 1, 0	0.0753	0, 2, 0	0.1349	1, 1, 1	0.1436	1, 1, 1
% Fear	0.0848	1, 1, 0	0.0759	2, 1, 0	0.1362	1, 1, 1	0.1370*	1, 1, 1
% Disgust	0.0857	1, 1, 0	0.0757	0, 2, 1	0.1339	1, 1, 1	0.1381	1, 1, 1
% Anticipation	0.0857	1, 1, 0	0.0752	0, 2, 1	0.1333	1, 1, 1	0.1421	1, 1, 1
% Anger	0.0856	1, 1, 0	0.0747*	1, 2, 0	0.1361	1, 1, 1	0.1383	1, 1, 1
Best Predictors (% Negative Sentiment, % Surprise, % Sadness)	0.0850	1, 1, 0	0.0752	2, 1, 0	0.1347	1, 1, 1	0.1413	1, 1, 1

Other metropolitan areas also showed improvements in vaccination predictions when Twitter features were added as exogenous variables. For example, in the Miami-Fort Lauderdale-Pompano Beach metropolitan area, adding the *number of users discussing the COVID-19 vaccine per 100,000 population* resulted in the best model and a lower RMSE compared to the baseline ARIMA model that did not contain exogenous features. In the Washington-Arlington-Alexandria metropolitan area, adding the *percentage of vaccine tweets expressing anger* or *joy* resulted in the best models and a lower RMSE compared to the baseline ARIMA model that did not contain any exogenous features. In the Chicago-Naperville-Elgin metro area, adding the *percentage of vaccine tweets with neutral sentiment* resulted in the best model and a lower RMSE compared to the baseline ARIMA model. In the Los Angeles-Long Beach-Anaheim metropolitan area, adding the *percentage of vaccine tweets expressing fear* resulted in the best model and a lower RMSE compared to the baseline ARIMA model.

Figure 5.2 illustrates the performance of the baseline ARIMA models (blue dotted line) and the best performing ARIMAX models (green dotted line), compared to the observed values of the outcome variable during the "out-sample" forecasting period (April 13th – May 20th, 2021) (red solid line). Across all metropolitan areas, the ARIMAX time series models with Twitter-derived features aligned more closely with the actual values of the vaccination rates compared to the baseline ARIMA model that relied on past historical vaccination data alone.

Figure 5.2 Predicted vs. Observed COVID-19 Vaccination Rates, January 1 – May 20th, 2021



5.4. Discussion

5.4.1. Principal Findings

In this study, I sought to determine whether supplementing forecast models with COVID-19 vaccine attitudes found in tweets – modeled via sentiments and emotions – improves over baseline models that only use historical vaccination data. When evaluating model performance across all metro areas, the addition of COVID-19 vaccine attitudes found in tweets resulted in improved model performance, as reflected by RMSE, when compared to baseline forecast models that did not include these features. Specifically, compared with the traditional ARIMA model with vaccination data alone, ARIMAX models with the predictions of both historical vaccination data and COVID-19 vaccine attitudes found in tweets reduced RMSE by as much as 9%.

5.4.2. Study findings in context

The ongoing COVID-19 pandemic emphasizes the need for innovative approaches to public health surveillance. The global public health community has monitored the COVID-19 pandemic by tracking case counts, hospitalizations, deaths, and vaccinations. For the United States, these datasets are publicly available. Forecasting case counts and vaccination rates using existing historical data has been a key approach in COVID-19 surveillance efforts [137]. Previous forecast models for predicting vaccine uptake rate relied on traditional ARIMA methods, where historical data was used to predict future rates [138]. However, social media data sources, such as Twitter, reveal society's attitudes towards the pandemic and current vaccination efforts on a real-time basis. This provides an opportunity for a large volume of raw and uncensored data related to vaccine attitudes, across various geographic locations, to be

leveraged for disease surveillance, which can subsequently be used to supplement and improve existing models.

The findings of the present study suggest that attitudes extracted from Twitter data can be added to existing forecast models for monitoring vaccination uptake across various metropolitan areas. In certain metropolitan areas, the mere volume of tweets and users engaged in vaccine-related conversations improved model performance when compared to baseline models. These results echo the findings in the study by [129], which revealed another social media source, Google Trends data, improved the prediction of COVID-19 vaccination uptake in Italy when compared to baseline models. In this study, Google Trends data was represented as the relative search volume for each vaccine-related keyword. Another similar study developed a framework for predicting vaccination rates in the United States based on traditional clinical data and web search queries [139]. The results of this study also revealed the ability for online networks to predict societal willingness to receive vaccinations. Specifically, the authors found a similar improvement in model performance as the present study – with a reduction in RMSE of 9.1%.

Although few studies sought to supplement current vaccine models with social media data, to our knowledge, there are no studies that go beyond the mere volume of relevant Twitter data and factor in the sentiment and emotion of vaccine-related conversations. Over the course of the pandemic, some states experienced low vaccination rates despite comprehensive vaccine roll out programs. In these cases, it is important to consider the public's emotions and sentiments towards vaccines. The present study contributes to the literature by evaluating the ability for sentiments and emotions related to the COVID-19 vaccine to predict

vaccine uptake. Specifically, the results show an improvement in model performance in certain metropolitan areas when models were supplemented with the percentage of tweets expressing anger, fear, joy, positive sentiment, or neutral sentiment. A study conducted by Alegado et al examined the association between sentiments and emotions found in tweets and vaccine uptake via regression coefficient analysis [137]. This study showed similar insights – tweets expressing fear, sadness, and anger appeared to be significantly associated with vaccination rates.

The results of the present study have several implications for the present COVID-19 response. Public health experts now argue that the traditional concept of herd immunity may not apply to COVID-19 [112]. Instead, the focus is to increase vaccination uptake to substantially control community spread, without the societal disruptions caused by the virus [113]. Accurately forecasting vaccination uptake allows policy makers and researchers to evaluate how close we are to achieving normalcy again. Additionally, similar algorithms allow public health practitioners to better anticipate vaccine uptake behaviors and therefore develop targeted policies. As the global community builds towards achieving herd immunity, researchers should also "listen" to the vaccine conversation on social media – monitoring misconceptions and misinformation and implementing targeted vaccine education campaigns that address these misconceptions. Although the COVID-19 pandemic appears to be improving, the present framework can also be used to improve vaccine forecast models for future pandemics.

5.4.3. Limitations and future work

It is important to note that the present study has some limitations. The study period was limited to the first half of 2021. However, vaccines were not yet available to most of the US adult population until April 2021. Therefore, the study period did not capture the height of vaccination efforts. Another limitation is that as the COVID-19 pandemic evolves, vaccine related keywords may change, requiring frequent updating of the model. Future work may involve the use of topic modeling to capture the general themes surrounding the COVID-19 pandemic.

Another limitation is related to the geographic scope of the present study. The present study only focused on forecasting vaccine uptake in the United States. However, it is important to note that vaccination efforts must be addressed on a global scale, not just domestically, for normalcy to be attained. Future work should consider collecting tweets and vaccination data from other countries to see if similar models improve vaccine forecasts globally. Additionally, the present study only examined tweets posted in the English language. This potentially excluded several relevant tweets related to the COVID-19 vaccine conversation. Future work should involve the use of sentiment and emotion classifiers that include lexicons in other languages.

5.4.4. Conclusion

Researchers have found that the internet and social media both play a role in shaping personal or parental choices about vaccinations. Although few previous studies have developed forecast models for COVID-19 vaccination rates in the United States, to our knowledge, there are no studies that aim to factor in the real-time vaccination attitudes present on Twitter. The

present study suggests the benefits of using the linguistic constructs found in tweets to improve predictions of the COVID-19 vaccination rate. In this study, I found that supplementing baseline forecast models with both historical vaccination data and COVID-19 vaccine attitudes found in tweets reduced RMSE by as much as 9%. Developing a predictive tool for vaccination uptake in the United States will empower public health researchers and decision makers to design targeted vaccination campaigns in hopes of achieving the vaccination threshold required for widespread population protection.

Chapter 6 – Conclusions, Future Work, and Limitations

Considering the limited resources across public health jurisdictions and high costs associated with traditional public health data sources, in this dissertation, I identify three ways social media data can be used as an alternative, viable and low-cost data source for public health research. In this dissertation, I present three studies that leverage social media as a data source, to answer research questions related to public health and compare traditional public health data sources to social media data sources. In the next few sections, I present a summary of findings and address study limitations, policy implications, and future directions.

6.1. Study 1: Using Social Media to Predict Food Deserts in the United States: Infodemiology Study of Tweets

Prior research has used data sources such as surveys, geographic information systems, and food store assessments to identify regions classified as food deserts, but these data sources can be costly and take a long time to collect. In this study, I introduced a novel approach to identifying food deserts in the United States using the linguistic constructs found in food-related tweets.

The results of this study revealed associations between a census tract being classified as a food desert and an increase in the number of tweets in a census tract that mentioned unhealthy foods (P=.03), including foods high in cholesterol (P=.02) or low in key nutrients such as potassium (P=.01). I also found an association between a census tract being classified as a food desert and an increase in the proportion of tweets that mentioned healthy foods (P=.03) and fast-food restaurants (P=.01) with positive sentiment. In addition, I found that including food ingestion language derived from tweets in classification models that predict food desert

status improves model performance compared with baseline models that only include socioeconomic characteristics.

These findings suggest that residents of census tracts unknowingly provide important information regarding their food environment on Twitter through the food ingestion language found in tweets. Therefore, social media data presents a more accessible, scalable, and cost-effective alternative to traditional public health data sources, such as surveys, GIS technology, and food store assessments, which can be costly and time consuming.

6.2. Study 2: Using COVID-19 Vaccine Attitudes Found in Tweets to Predict Vaccine Perceptions in Traditional Surveys: Infodemiology Study of Tweets

Traditionally, surveys are conducted to answer questions related to public health but can be costly to execute. However, the information that researchers aim to extract from surveys could be potentially retrieved from social media – data that is highly accessible and lower in cost to collect. In this study, I evaluated whether attitudes towards COVID-19 vaccines collected from the Household Pulse Survey can be predicted using attitudes extracted from Twitter. Ultimately, I sought to determine whether Twitter can provide us with similar information to what is observed in traditional surveys, or, if saving money comes at the cost of losing rich data.

The results of this study revealed that attitudes toward COVID-19 vaccines found in tweets explained 61-72% of the variability in the percentage of HPS respondents that were vaccine hesitant or compliant. I also found significant statistical relationships between perceptions expressed on Twitter and in the survey. A linear regression analysis showed significant relationships between (1) the percentage of vaccine compliant HPS respondents and the percentage of tweets expressing positive sentiment, joy, trust, and anticipation; and (2) the

percentage of vaccine hesitant HPS respondents and the percentage of tweets expressing negative sentiment, anger, and anticipation.

These findings suggest the information researchers aim to extract from surveys could also potentially be retrieved from a more accessible and cost-effective data source, such as Twitter data. The results of this study support the efforts of researchers, who over the past few years have looked to social media as a data source instead of traditional surveys, citing the availability of more readily available data and no or low-cost data collection efforts.

6.3. Study 3: Using COVID-19 vaccine tweets to predict vaccination rates in the United States: Infodemiology Study of Tweets

One of the most common forecasting methods used, univariate time series, involves predicting future vaccination rates using historical vaccination rates. While this method can be useful in many cases, it fails to account for other time-dependent factors that may also influence vaccinations. Researchers have found that the internet and social media both play a role in shaping personal choices about vaccinations. Therefore, it may be important to consider the daily conversations on social media when developing vaccine uptake forecast models. The goal of this study was to determine whether supplementing COVID-19 vaccine uptake forecast models with the attitudes found in tweets improves over baseline models that only use historical vaccination data.

When evaluating model performance across all metro areas, the addition of COVID-19 vaccine attitudes found in tweets resulted in improved model performance, as reflected by RMSE, when compared to baseline forecast models that did not include these features. Specifically, compared with the traditional ARIMA model with vaccination data alone, ARIMAX

models with the predictions of both historical vaccination data and COVID-19 vaccine attitudes found in tweets reduced RMSE by as much as 9%.

Developing a predictive tool for vaccination uptake in the United States will empower public health researchers and decision makers to design targeted vaccination campaigns in hopes of achieving the vaccination threshold required for the United States to reach widespread population protection.

6.4. Policy Implications and Future Directions

This dissertation has one overarching public health implication: Twitter data can be used as a viable data source for public health research – a data source that has been deemed attractive due the geographic granularity of such novel information, and importantly, the speed of data collection. For researchers, this dissertation highlights the importance of evaluating ways to modernize public health data and methods, especially considering the public health threats of the 21st century, such as COVID-19, natural disasters, drug abuse, and mental health. Traditional data sources used in public health research come with their own challenges, but most notably the expensive administration costs and low response rates associated with surveys; the delay and sometimes sparsity of public health surveillance data; and the limited reach and availability of medical records. Considering the chronic underfunding of public health jurisdictions in the United States, it is even more important for public health jurisdictions to find effective and low-cost ways to address the increasing public health challenges of the 21st century.

Why is Twitter an attractive data source compared to more traditional public health data sources? First, we have the speed of data collection. The free, publicly available Twitter

API makes it possible for users to develop software that integrates Twitter, as well as provides access to public Twitter data, both in real-time and in the past. Academic researchers make up a large proportion of users who take advantage of the Twitter API, using the various Twitter APIs to collect and analyze the public conversations found in tweets [47]. Next, we have the costs associated with data collection, or lack thereof. Data used in this dissertation were collected at no cost via the Twitter Streaming API. There is also the geographic granularity of tweets. As shown in the three studies presented in this dissertation, many tweets are tagged with some type of location information, allowing researchers to answer research questions specific to various geographic areas. Finally, social media companies such as Twitter acknowledged that academic researchers are one of the largest groups of people using their APIs, and have taken the necessary steps to improve the experience for researchers and therefore facilitate the use of social media data to advance our disciplines [47].

Future work can be done to expand the impact of the studies presented in this dissertation. Each of the studies presented in this dissertation focused on large metropolitan areas in the United States and only examined tweets posted in the English language. However, it is important to note that public health issues such as food insecurity, vaccine hesitancy, and vaccine uptake should also be addressed on a global scale, not just domestically. Future work should consider the application of similar studies on a global scale, including more rural areas that are not typically the focus of large-scale public health efforts. Future work should involve the use of sentiment and emotion classifiers that include lexicons in other languages. Future work should also involve the improvement of the accuracy of sentiment and emotion classifiers.

This dissertation focuses on social media, specifically Twitter data, as a data source for public health research. However, social media has the potential to play other roles in public health. For example, social media can be used as a tool to increase the impact of public health research by providing avenues of disseminating research, combatting misinformation, influencing policy, and enhancing professional development of public health practitioners [140]. Future work should examine the ways in which social media allows public health researchers to shape messaging and public discourse, disseminate their work, and affect policy [141].

6.5. Limitations and Privacy Concerns

Although each of the three studies in this dissertation explored different uses of Twitter data, there were common limitations across all the studies. A key limitation of Twitter data is lack of representation among Twitter users, which suggests limited generalizability of results to a larger population. In general, Twitter users tend to be younger, more educated, have higher incomes, and are more liberal [16]. Skewed age in the tweets sample can potentially impact findings, especially when exploring food consumption habits or perceptions towards vaccines. For example, a study conducted by Allman-Farinelli et al. showed that young people preferred and overconsumed unhealthy foods and favored some food products more than older adults, such as alcoholic beverages and foods high in sugar [142]. These food consumption habits may also appear dominant on Twitter due to the younger user demographic. Demographics such as age are not widely available for Twitter users, making it difficult to quantify the impact of this limitation.

The disparity in Twitter activity among Twitter users can also be noted as a limitation of Twitter data. Twitter users tweet a median of only 2 tweets per month, with just 10% of Twitter users accounting for 80% of the tweets across users in the United States [44]. This disparity in Twitter activity suggests that a large sample of tweets may actually reflect a much smaller sample of individuals. Despite these limitations, in all three studies, I obtained results that would suggest that the tweets collected contained a significant signal for conducting public health research. For example, (1) supplementing classification models with features derived from food ingestion language found in tweets, such as positive sentiment toward mentions of healthy foods and fast-food restaurants, improved baseline models that only included demographic and SES features by up to 19%, with AUC scores >0.8; (2) the attitudes toward COVID-19 vaccines found in tweets explained 61-72% of the variability in the percentage of HPS respondents that were vaccine hesitant or compliant; and (3) supplementing baseline vaccine uptake forecast models with both historical vaccination data and COVID-19 vaccine attitudes found in tweets improved forecast models and reduced RMSE by as much as 9%.

Another limitation to note is the mere volume of tweets available via the Twitter Streaming API was limited to a random sample of 1% of all Tweets sent by Twitter users at any given time [14]. Each of the studies presented in this dissertation include some level of geographic granularity, which required tweets to have geolocation information. However, studies have shown that only approximately 1% to 2% of the tweets from the Twitter streaming API include geolocation information [20]. Although this was a limitation that I experienced during my data collection period, academic researchers now have free access to the entire historical archive of public Twitter data via the Academic Track Twitter API.

With social media being increasingly used to conduct research, it is also important to note the privacy concerns associated with the use of social media for public health research. Social media platforms such as Twitter provide free access to user-generated social media content. Just like other types of research involving human subjects, researchers should comply with the highest standards of data privacy and data protection, even if social media data are considered publicly available [143]. Misuses of social media data may lead to individual users' privacy being compromised [143]. Since it is virtually impossible to obtain consent from all social media users included in a sample, it is important that all content that can be used to identify social media users be removed in order to protect the privacy of the users whose data we collect [144]. This protection is especially important when conducting research on sensitive topics, such as drug abuse, mental health, politics, and reproductive rights. It is important to note that the studies in this dissertation used fully aggregated Twitter data – therefore mitigating potential privacy violations. Additionally, the University of Maryland College Park IRB determined these studies did not meet the definition of human subject research under the purview of the IRB according to federal regulations.

Appendix A: Food Keyword List

good times burgers & frozen custard
mint chocolate chip ice cream
dominos philly cheese steak pizza
pizza hut stuffed crust pizza
bojangles famous chicken n biscuits
freddys frozen custard & steakburgers
penn station east coast subs
michelob ultra lime cactus
miller high life light
shock top raspberry wheat
black and white cookie
cream cheese with herbs
ham and cheese sandwich
double whopper with cheese
premium alaskan fish sandwich
butter pecan ice cream
chocolate chip ice cream
cookie dough ice cream
french vanilla ice cream
rocky road ice cream
new york strip steak
pork baby back ribs
peanut butter toast crunch
special k chocolatey delight
special k protein plus
special k red berries
new york style pizza
pizza hut supreme pizza
canada dry ginger ale
chicken with rice soup
cream of asparagus soup
cream of broccoli soup
cream of celery soup
cream of chicken soup
cream of mushroom soup
cream of onion soup
cream of potato soup
creamy chicken noodle soup
yoplait boston cream pie
yoplait key lime pie
captain ds seafood kitchen
dixie chili and deli
green burrito red burrito
the habit burger grill
jack in the box
lees famous recipe chicken
raising canes chicken fingers
cranberry apple juice
cranberry grape juice
passion fruit juice
pineapple orange juice
white grape juice
baileys irish cream
beef minute steak
beef prime rib
rib eye steak
veal roast beef
bud light chelada
bud light lime
bud select 55
genesee cream ale
hurricane high gravity
michelob amber bock
michelob ultra amber
miller genuine draft
miller high life
milwaukees best light
old milwaukee na
pabst blue ribbon
rolling rock light
sierra nevada strong
st pauli girl
angel food cake
black forest cake
chocolate cream pie
chocolate mousse cake
chocolate mousse pie
flourless chocolate cake
german chocolate cake
ice cream cake
key lime pie
lemon meringue pie
pineapple upsidedown cake
red velvet cake
strawberry rhubarb pie
sweet potato pie
tres leches cake
victoria sponge cake
mike and ike
peanut butter bars
peanut butter cookies
sour patch kids
canned crushed pineapple
canned fruit cocktail
canned fruit salad
canned mandarin oranges
canned mixed fruit
canned morello cherries
canned sliced pineapple
canned sour cherries
durum wheat semolina
whole grain wheat

monterey jack cheese
chicken breast fillet
ovenroasted turkey breast
smoked turkey breast
chive cream cheese
feta cream cheese
garlic cream cheese
olive cream cheese
pesto cream cheese
philadelphia cream cheese
pineapple cream cheese
salmon cream cheese
vegetable cream cheese
walnut cream cheese
baby back ribs
bacon and eggs
chicken caesar salad
chicken fried steak
chicken pot pie
chicken tikka masala
chili con carne
corned beef hash
fish and chips
grilled cheese sandwich
mac and cheese
macaroni and cheese
peanut butter sandwich
philly cheese steak
pulled pork sandwich
grand turkey club
roast beef classic
roast beef max
big n tasty
original chicken sandwich
whopper with cheese
chicken teriyaki sandwich
grilled chicken salad
filet o fish
nachos with cheese
jr bacon cheeseburger
son of baconator
ben and jerrys
chocolate ice cream
coffee ice cream
cold stone creamery
dairy milk mcflurry
hot fudge sundae
ice cream sandwich
ice cream sundae
magnum double caramel
magnum double chocolate
snickers ice cream
strawberry ice cream
vanilla ice cream
fried bean curd
red kidney bean
soy nut butter
textured soy protein
textured vegetable protein
flat iron steak
pork countrystyle ribs
standing rib roast
sweetened condensed milk
cinnamon toast crunch
cracklin oat bran
cream of wheat
crunchy nut cornflakes
honey nut cheerios
kelloggs corn flakes
post raisin bran
post shredded wheat
quaker oatmeal squares
raisin bran crunch
raisin nut bran
low carb pasta
whole grain noodles
whole grain spaghetti
banana nut bread
conchasmexican sweet bread
hot dog buns
oatmeal raisin cookies
pan de sal
whole wheat bread
bbq chicken pizza
buffalo chicken pizza
deep dish pizza
four cheese pizza
goat cheese pizza
quattro formaggi pizza
red pepper pizza
spinach feta pizza
stuffed crust pizza
thin crust pizza
pork blade steak
pork crown roast
pork rib roast
pork shoulder blade
rack of pork
french fingerling potatoes
japanese sweet potatoes
norland red potatoes
potatoes au gratin
purple majesty potatoes
red gold potatoes
russian banana potatoes
yukon gold potatoes
breakfast sausage links
cold pack cheese
lacy swiss cheese
diet cherry coke
diet dr pepper
minute maid light
mug root beer
schweppes ginger ale
beef noodle soup
broccoli cheese soup
carrot ginger soup
chicken gumbo soup
chicken noodle soup
chicken vegetable soup

french onion soup
golden mushroom soup
lobster bisque soup
tomato rice soup
vegetable beef soup
mallard duck meat
wild boar meat
dry red wine
sweet red wine
sweet white wine
yellow tail wine
aloe vera yogurt
bircher muesli yogurt
skim milk yogurt
yoplait french vanilla
yoplait greek blueberry
yoplait greek coconut
yoplait greek strawberry
yoplait greek vanilla
yoplait harvest peach
yoplait mixed berry
yoplait pina colada
yoplait strawberry banana
yoplait strawberry cheesecake
arctic circle restaurants
atlanta bread company
au bon pain
buffalo wild wings
california pizza kitchen
carinos italian grill
charleys philly steaks
chick fil a
chuck a rama
chuck e cheese
el pollo loco
el taco tote
gold star chili
in n out
jersey mikes subs
l&l hawaiian barbecue
lee roy selmons
long john silvers
noodles and company
port of subs
roy rogers restaurants
seattles best coffee
steak n shake
tudors biscuit world
beef ribs
acai juice
aloe vera
apple juice
apricot nectar
banana juice
blackberry juice
boysenberry juice
carrot juice
chamomile tea
cherry juice
coconut milk
coconut water
concord grape
cranberry juice
cucumber juice
currant juice
grape juice
grapefruit juice
lemon juice
lime juice
mango lassi
orange juice
papaya juice
peach juice
peach nectar
pear juice
pear nectar
pineapple juice
plum juice
pomegranate juice
sauerkraut juice
tangerine juice
tomato juice
vegetable juice
blue curacao
canadian whiskey
grand marnier
irish whiskey
jack daniels
jim beam
red wine
rose wine
sloe gin
southern comfort
triple sec
white wine
beef brisket
beef fillet
beef goulash
beef neck
beef pancreas
beef patty
beef ribs
beef sirloin
beef suet
beef tallow
beef thymus
beef tripe
chuck roast
chuck steak
filet mignon
flank steak
ground beef
ground chuck
ground round
minced veal
porterhouse steak
roast beef
rump steak
skirt steak
stew beef

strip steak
veal breast
veal leg
veal shank
veal shoulder
veal sirloin
veal tenderloin
blue moon
bock beer
buckler na
bud ice
bud light
bud select
budweiser chelada
busch light
busch na
clausthaler na
colt 45
coors light
coors na
dark beer
ginger beer
honey brown
keystone ice
keystone light
lager beer
land shark
light beer
malt beer
michelob lager
michelob light
michelob ultra
miller chill
miller lite
milwaukees best
natural ice
natural light
nonalcoholic beer
odouls na
old milwaukee
olde english
pale ale
redbridge glutenfree
rolling rock
root beer
samuel adams
shock top
steel reserve
stout beer
strong beer
wheat beer
apple cake
apple cobbler
apple crisp
apple crumble
apple pie
apple strudel
apple turnover
applesauce cake
baked alaska
bakewell tart
banoffee pie
birthday cake
blueberry cobbler
blueberry muffin
blueberry pie
bundt cake
buttermilk pie
caramel cake
carrot cake
cherry pie
chess pie
chocolate cake
chocolate muffin
coconut cake
coffee cake
cream puff
crumb cake
danish pastry
french cruller
fruit cake
funnel cake
king cake
layer cake
lemon cake
madeira cake
marble cake
opera cake
peach cobbler
peach pie
pecan pie
plum cake
poppyseed cake
pound cake
puff pastry
pumpkin bread
pumpkin cheesecake
pumpkin pie
raspberry pie
rhubarb pie
rum cake
sacher torte
sponge cake
strawberry cheesecake
strawberry pie
swiss roll
tarte tatin
tiramisu cake
treacle tart
wedding cake
3 musketeers
after eight
almond roca
angel delight
animal crackers
baby ruth
buttermilk pancakes
candy apple
candy cane
candy canes

candy corn
candy floss
caramel popcorn
caramel squares
chocolate bar
chocolate chips
cotton candy
eggy bread
fairy cakes
ferrero rocher
fortune cookies
granola bars
gummi bears
hershey kisses
jelly beans
jelly belly
jolly ranchers
jordan almonds
kit kat
laffy taffy
lindt chocolate
mars bar
milk duds
milky way
peanut bar
peanut brittle
peppermint bark
pop rocks
pumpkin seeds
rice pudding
roasted almonds
spritz cookies
take 5
whoopie pie
canned apricots
canned blackberries
canned blueberries
canned cherries
canned cranberries
canned figs
canned gooseberries
canned grapefruit
canned grapes
canned mango
canned mangosteen
canned oranges
canned peaches
canned pears
canned pineapple
canned plums
canned raspberries
canned strawberries
canned tangerines
dried fruit
barley groats
brown rice
buckwheat groats
corn waffles
millet flour
millet gruel
oat bran
pearl barley
prawn crackers
pretzel sticks
rye bran
savoury biscuits
spelt bran
spelt semolina
sunflower seeds
tortilla chips
wheat bran
wheat germ
wheat gluten
wheat semolina
wheat starch
wholegrain oat
american cheese
asiago cheese
blue cheese
cheese fondue
cheese spread
cheese whiz
colby cheese
colbyjack cheese
cottage cheese
dutch cheese
edam cheese
fresh mozzarella
goat cheese
grated parmesan
grilled cheese
italian cheese
maasdam cheese
manchego cheese
muenster cheese
raclette cheese
sheep cheese
soft cheese
stilton cheese
string cheese
swiss cheese
white cheddar
wisconsin cheese
baked ham
beef salami
boiled ham
chopped ham
corned beef
dutch loaf
ham sausage
head cheese
hickory ham
honey ham
olive loaf
parma ham
pimento loaf
pork roast
serrano ham
smoked ham
summer sausage

turkey breast
turkey ham
turkey salami
cream cheese
baked beans
baked chicken
bbq ribs
beef stew
black pudding
black rice
butter chicken
california roll
chicken marsala
chicken parmesan
cobb salad
collard greens
corn dog
cottage pie
deviled eggs
dim sum
fried rice
fried shrimp
mashed potatoes
meat pie
orange chicken
pad thai
pea soup
peking duck
pork chop
potato salad
reuben sandwich
roast dinner
sausage roll
sausage rolls
shepherds pie
shrimp cocktail
sloppy joe
sloppy joes
spaghetti bolognese
spring roll
spring rolls
tandoori chicken
yorkshire pudding
bbq rib
bean burrito
angry whopper
double whopper
triple whopper
whopper jr
chicken breast
chicken fajita
chicken mcnuggets
chicken nuggets
chicken pizziola
chicken sandwich
chicken wings
chop suey
curly fries
double cheeseburger
egg roll
fish sandwich
french fries
ham sandwich
hot dog
italian bmt
big mac
egg mcmuffin
mighty wings
meatball sandwich
onion rings
smoked salmon
spicy italian
club sandwich
tortilla wrap
veggie burger
veggie delight
veggie patty
jr cheeseburger
zinger burger
fish fingers
fish sticks
pickled herring
red snapper
sea bass
tuna salad
blood oranges
custard apple
fruit salad
mandarin oranges
passion fruit
ciao bella
crunchie mcflurry
dippin dots
double rainbow
healthy choice
ice milk
magnum almond
magnum gold
magnum white
mcflurry oreo
mini milk
smarties mcflurry
soft serve
strawberry sundae
turkey hill
vanilla cone
azuki bean
bamboo shoots
bean curd
bean sprouts
bengal gram
black beans
black chickpeas
black gram
brown lentil
chili bean
deepfried tofu
extrafirm tofu
firm tofu
green beans

green gram
green lentil
kidney beans
lima beans
mung beans
navy bean
pinto beans
puy lentils
red beans
red lentils
refried beans
roasted soybeans
silken tofu
soft tofu
soy cheese
soy mayonnaise
soy nuts
soy yogurt
soya cheese
soynut butter
sugar peas
sweet peas
white beans
yam bean
yellow lentils
beef jerky
beef tenderloin
chicken drumstick
chicken fat
chicken giblets
chicken gizzards
chicken leg
chicken liver
chicken meat
chicken thigh
chicken wing
cubed steak
pork chops
pork loin
pork steaks
round steak
spare ribs
tbone steak
turkey legs
turkey steak
turkey wings
almond milk
chocolate mousse
coffee creamer
condensed milk
creme fraiche
evaporated milk
goat milk
hot chocolate
lactosefree milk
plain yogurt
powdered milk
rice milk
semiskimmed milk
semolina pudding
skim milk
sour cream
soy milk
whipped cream
whole milk
ace drink
apple spritzer
chai tea
chocolate milk
chocolate milkshake
club mate
coca cola
coke zero
crystal light
diet coke
egg cream
egg nog
elderflower cordial
ginger tea
hawaiian punch
ice tea
iced tea
latte macchiato
milkshake dry
slush puppie
strawberry milkshake
tap water
vanilla milkshake
yerba mate
alfalfa sprouts
brazil nuts
chia seeds
cotton seeds
ginkgo nuts
goa bean
hickory nuts
lotus seed
macadamia nuts
pecan nuts
pili nuts
pine nuts
poppy seeds
radish seeds
safflower seeds
sesame seeds
smoked almonds
soy beans
sweet chestnut
bran flakes
capn crunch
chocolate cheerios
coco pops
cocoa krispies
cocoa pebbles
cocoa puffs
cookie crisp
corn flakes
corn pops
count chocula
crunchy nut

fiber one
froot loops
frosted cheerios
frosted flakes
frosted miniwheats
fruity pebbles
golden grahams
honey smacks
just right
life cereal
lucky charms
multigrain cheerios
puffed rice
puffed wheat
quaker grits
quaker oatmeal
raisin bran
ready brek
rice chex
rice krispies
shredded wheat
smart start
special k
sugar puffs
cellophane noodles
cheese tortellini
dumpling dough
egg noodles
glass noodles
lasagne sheets
penne rigate
shirataki noodles
soy noodles
spinach tortellini
banana bread
beer bread
black bread
bran muffins
bread pudding
brown bread
cinnamon bun
english muffin
garlic bread
italian bread
matzo bread
monkey bread
multigrain bread
oatmeal cookies
pita bread
potato bread
pretzel roll
raisin bread
rye bread
sandwich bread
shortcrust pastry
soda bread
sourdough bread
spice cake
sweet rolls
tortilla bread
white bread
bbq pizza
beef pizza
bianca pizza
calabrese pizza
capricciosa pizza
cheese pizza
chicken pizza
grilled pizza
hawaiian pizza
margherita pizza
mozzarella pizza
mushroom pizza
napoli pizza
pepperoni pizza
pizza dough
pizza rolls
salami pizza
sausage pizza
seafood pizza
shrimp pizza
sicilian pizza
spinach pizza
tarte flambe
tuna pizza
vegetable pizza
vegetarian pizza
veggie pizza
white pizza
canadianstyle bacon
ground pork
hog maws
pig ear
pig fat
pigs tail
pigs trotter
pork bacon
pork belly
pork cutlet
pork cutlets
pork leg
pork meatloaf
pork ragout
pork ribs
pork sausage
pork shank
pork shoulder
pork stomach
salt meat
salt pork
allblue potatoes
baked potato
boiled potatoes
fried potatoes
marrow dumplings
potato dumpling
potato fritter
potato gratin
potato pancakes
potato starch

potato sticks
potato waffles
potato wedges
red potatoes
roast potatoes
rosemary potatoes
russet potatoes
sweet potato
white potatoes
chicken drumsticks
chicken legs
chicken thighs
cornish hens
duck breast
ostrich meat
pheasant breast
pheasant leg
quail breast
turkey cutlet
turkey drumsticks
wild duck
blood sausage
cheese pastry
chicken salad
cooked ham
cumberland sausage
garlic sausage
hot dogs
hot sausage
italian sausage
liver pate
luncheon meat
polish sausage
pork roll
ring bologna
smoked sausage
butter cheese
cheese platter
cheese slices
esrom cheese
fol epi
gouda cheese
leerdammer cheese
pepper cheese
sandwich cheese
smoked cheese
white american
young gouda
bitter lemon
cherry coke
diet pepsi
diet sunkist
dr browns
dr pepper
fanta zero
five alive
full throttle
fuze tea
ginger ale
jolt cola
mello yello
minute maid
orange soda
pibb xtra
sprite zero
tonic water
vanilla coke
alphabet soup
bean stew
beef bouillon
beef soup
broccoli soup
cabbage soup
carrot soup
chicken bouillon
chicken broth
chicken stock
instant ramen
lentil soup
meatball soup
mushroom soup
noodle soup
onion soup
oxtail soup
potato soup
pumpkin soup
scotch broth
thai soup
tomato soup
vegetable broth
vegetable soup
vegetable stock
wedding soup
asian pear
cantaloupe melon
casaba melon
dragon fruit
galia melon
maraschino cherries
pink grapefruit
prickly pear
soursop fruit
star fruit
bell pepper
black olives
brussels sprouts
cherry tomato
chinese cabbage
creamed spinach
green olives
green onion
mustard greens
red cabbage
turnip greens
winter squash
alligator meat
antelope meat
bear meat
beaver meat
bison meat

bison sirloin
buffalo meat
goose meat
lamb meat
mallard meat
moose meat
mutton meat
pheasant meat
racoon meat
reindeer meat
squirrel meat
venison sirloin
wild boar
cabernet sauvignon
chenin blanc
chocolate wine
malbec wine
marsala wine
merlot wine
moscato wine
mulled wine
pinot gris
pinot noir
plum wine
port wine
red wines
riesling wine
sauvignon blanc
sparkling wine
sweet wines
white zinfandel
activia lemon
activia strawberry
blueberry yogurt
cherry yogurt
chocolate yogurt
cream yogurt
creamy yogurt
diet yogurt
fruit yogurt
greek yogurt
lowfat yogurt
mocca yogurt
organic yogurt
peach yogurt
probiotic yogurt
stracciatella yogurt
strawberry yogurt
vanilla yogurt
yogurt corner
yoplait mango
yoplait strawberry
baskin robbins
a&w restaurants
arthur treachers
auntie annes
baja fresh
boston market
burger chef
burger king
burger street
carls jr
cheeburger cheeburger
chicken express
churchs chicken
cocos bakery
cold stone
cook out
dairy queen
del taco
duck donuts
dunkin donuts
el chico
elevation burger
farmer boys
firehouse subs
five guys
fosters freeze
golden chick
halal guys
huddle house
jamba juice
jasons deli
jimmy johns
jims restaurants
johnny rockets
krispy kreme
lions choice
little caesars
maid rite
marcos pizza
mcalisters deli
milos hamburgers
mod pizza
mrs fields
mrs winners
panda express
papa johns
papa murphys
pita pit
pizza hut
pollo tropical
shake shack
skyline chili
sneaky petes
steak escape
taco bell
taco bueno
taco cabana
taco johns
taco mayo
taco tico
taco time
tim hortons
twin peaks
umami burger
wetzels pretzels
white castle
caprisun
clamato

lemonade
limeade
nestea
absinthe
amaretto
applejack
asti
baileys
beer
bourbon
brandy
chambord
champagne
cider
cognac
cointreau
drambuie
frangelico
gin
jagermeister
kahlua
liqueur
liquor
midori
prosecco
punch
rum
sambuca
sangria
scotch
sherry
tequila
vermouth
vodka
whisky
wine
beef
veal
ale
bass
becks
budweiser
busch
coors
corona
guinness
heineken
ipa
molson
pilsner
porter
sparks
stout
tilt
cheesecake
cupcake
donut
doughnut
flan
gingerbread
meringue
paczki
pancake
panettone
profiterole
tiramisu
trifle
waffles
airheads
butterfinger
celebrations
chocolate
cookies
flapjack
gumdrops
licorice
lifesavers
liquorice
lollipop
mms
maltesers
marshmallows
marzipan
payday
pez
popcorn
rolo
smores
semolina
skittles
smarties
snickers
speculoos
toblerone
twix
applesauce
amaranth
barley
buckwheat
cornmeal
cornstarch
couscous
cracker
flaxseed
freekeh
gluten
grissini
kamut
millet
polenta
quinoa
rusk
sago
shortbread
spelt
tortilla
applewood
babybel
brie
camembert

cheddar
chester
emmentaler
feta
fontina
gjetost
gorgonzola
gouda
gruyere
halloumi
havarti
jarlsberg
monterey
mozzarella
neufchatel
parmesan
pecorino
provolone
ricotta
romano
roquefort
bologna
bresaola
capicola
chorizo
ham
liverwurst
pastrami
pepperoni
prosciutto
salami
tongue
horseradish
mascarpone
biryani
blt
burrito
chimichanga
dal
dosa
enchiladas
fajita
hummus
jambalaya
kebab
lasagne
meatloaf
naan
paella
paratha
pizza
ramen
ravioli
samosa
taco
reuben
bratwurst
whopper
cheeseburger
falafel
hamburger
lasagna
mcchicken
mcdouble
mcrib
tuna
turkey
baconator
zinger
poutine
anchovy
bluefish
bream
butterfish
calamari
carp
caviar
clam
cod
crawfish
crayfish
eel
flounder
grouper
haddock
hake
halibut
herring
hoki
kipper
ling
lobster
mackerel
milkfish
monkfish
mullet
mussel
octopus
pickerel
pike
pollack
redfish
salmon
sardines
scallops
scampi
shad
shark
smelt
sole
squid
sturgeon
sushi
swordfish
trout
turbot
wahoo
whitefish
whiting
acai

apple
apricot
avocado
banana
blackberries
blueberries
cantaloupe
cherries
clementine
cranberries
currants
dates
figs
grapes
guava
jackfruit
jujube
kiwi
lemon
lime
lychees
mango
mulberries
nectarine
olives
orange
papaya
peach
pear
persimmon
physalis
pineapple
plantains
plum
pomegranate
quince
raisins
rambutan
raspberries
rhubarb
starfruit
strawberries
tamarind
tangerine
watermelon
carvel
drumsticks
friendlys
magnolia
magnum
mcflurry
schwans
solero
sundae
butternut
chickpeas
flageolet
lentils
marron
miso
natto
okara
okra
peanuts
peas
pecan
rajma
soybeans
tempeh
tofu
yuba
alligator
chicken
duck
ostrich
pork
schnitzel
buttermilk
cream
curd
custard
kefir
lassi
milk
quark
tzatziki
yogurt
chai
coffee
cola
evian
gatorade
hic
horchata
kombucha
milo
nectar
powerade
slurpee
smoothie
tang
tea
water
acorn
almonds
beechnut
breadfruit
cashew
chestnut
coconut
hazelnut
pecans
pistachios
walnuts
cheerios
chex
chocos
frosties
honeycomb
kashi

kix
krave
muesli
oatmeal
puff
scooters
toasties
trix
weetabix
wheaties
cannelloni
capellini
farfalle
fettuccine
fusilli
linguine
macaroni
manicotti
mostaccioli
orecchiette
orzo
penne
pierogi
rigatoni
rotini
shells
spaetzle
spaghetti
tagliatelle
tortellini
vermicelli
ziti
bagel
baguette
biscuit
breadsticks
brioche
brownies
bun
cannoli
challah
chapati
ciabatta
cornbread
crepes
croissant
crumpet
cupcakes
empanada
flatbread
focaccia
latkes
muffin
pie
pretzel
pumpernickel
roll
roti
sandwich
scone
souffl
spanakopita
toast
calzone
bacon
chitterlings
lard
spareribs
cassava
croquettes
dumplings
gnocchi
omelette
potato
rsti
yams
capon
dove
emu
goose
pheasant
pigeon
poularde
quail
andouille
bockwurst
boudin
frankfurters
jerky
kielbasa
knackwurst
landjaeger
linguica
mettwurst
mortadella
sausage
souse
spam
weisswurst
edam
emmental
tilsit
7up
barqs
coke
fanta
fresca
fuze
pepsi
soda
sprite
squirt
sunkist
surge
tab
bouillon
goulash
minestrone
succotash
acerola

cherimoya
durian
feijoa
grapefruit
honeydew
kumquat
lychee
mangosteen
maracuya
muskmelon
noni
plantain
pomelo
sapodilla
artichoke
arugula
asparagus
aubergine
beetroot
broccoli
cabbage
capsicum
carrot
cauliflower
celery
chard
chicory
chives
corn
courgette
cucumber
eggplant
endive
fennel
garlic
gherkin
gourd
kale
kohlrabi
leek
lettuce
mushrooms
nori
onion
parsnips
pepper
pumpkin
radishes
rutabaga
shallots
spinach
squash
tomato
turnips
wasabi
zucchini
bison
buffalo
caribou
lamb
moose
mutton
rabbit
reindeer
squirrel
venison
cava
chardonnay
gamay
merlot
riesling
shiraz
zinfandel
activia
ayran
arbys
blimpie
bojangles
braums
burgerville
checkers
rallys
chicfila
chipotle
chuckarama
cinnabon
culvers
dibellas
dominos
druthers
duchess
eegees
fatburger
hardees
kfc
kewpee
krystal
mcdonalds
moes
mooyah
naugles
panera
popeyes
potbelly
qdoba
quiznos
rax
robeks
saladworks
sbarro
schlotzskys
smashburger
sonic
spangles
starbucks
subway
swensens
swensons
wendys
whataburger

wienerschnitzel	wingstop	zippys
wimpy	zaxbys	

Appendix B: COVID-19 Vaccine Keyword List

phizer
vaccine
moderna
dose
doses
vaccines
johnson and
johnson
johnson johnson
johnson & johnson
johnsonandjohnson
johnson&johnson
johnsonjohnson
shot

Bibliography

[1] Institute of Medicine (US) Committee on Health Research and the Privacy of Health Information: The HIPAA Privacy Rule, *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. Washington (DC): National Academies Press (US), 2009. Accessed: Jan. 13, 2022. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK9578/

[2] "Health Data Sources." https://www.nlm.nih.gov/nichsr/stats_tutorial/section3/index.html (accessed Jan. 13, 2022).

[3] "Surveys." https://www.nlm.nih.gov/nichsr/stats_tutorial/section3/mod1_surveys.html (accessed Jan. 13, 2022).

[4] "Surveillance." https://www.nlm.nih.gov/nichsr/stats_tutorial/section3/mod5_surveillance.html (accessed Jan. 13, 2022).

[5] "Medical Records." https://www.nlm.nih.gov/nichsr/stats_tutorial/section3/mod2_medical.html (accessed Jan. 13, 2022).

[6] E. L. Baker *et al.*, "The public health infrastructure and our nation's health," *Annu Rev Public Health*, vol. 26, pp. 303–318, 2005, doi: 10.1146/annurev.publhealth.26.021304.144647.

[7] 1615 L. St NW, S. 800 Washington, and D. 20036 U.-419-4300 | M.-857-8562 | F.-419-4372 | M. Inquiries, "Demographics of Social Media Users and Adoption in the United States," *Pew Research Center: Internet, Science & Tech*. https://www.pewresearch.org/internet/fact-sheet/social-media/ (accessed Jan. 13, 2022).

[8] H. Gu *et al.*, "Importance of Internet Surveillance in Public Health Emergency Control and Prevention: Evidence From a Digital Epidemiologic Study During Avian Influenza A H7N9 Outbreaks," *J Med Internet Res*, vol. 16, no. 1, p. e20, Jan. 2014, doi: 10.2196/jmir.2911.

[9] K. S. Sahu, S. E. Majowicz, J. A. Dubin, and P. P. Morita, "NextGen Public Health Surveillance and the Internet of Things (IoT)," *Front Public Health*, vol. 9, p. 756675, Dec. 2021, doi: 10.3389/fpubh.2021.756675.

[10] *USDA ERS - About the Atlas, 2019*. [Online]. Available: https://www.ers.usda.gov/data-products/food-access-research-atlas/about-the-atlas/#definitions.

[11] M. D. Choudhury, M. Gamon, S. Counts, and E. Horvitz, *Predicting Depression via Social Media*. 2013.

[12] M. J. Paul, M. Dredze, and J. H. University, *You Are What You Tweet: Analyzing Twitter for Public Health*. 2011.

[13] A. Sadilek, H. Kautz, and V. Silenzio, *Modeling Spread of Disease from Social Interactions*. 2012.

[14] Q. C. Nguyen *et al.*, "Building a National Neighborhood Dataset From Geotagged Twitter Data for Indicators of Happiness, Diet, and Physical Activity," *JMIR Public Health and Surveillance*, vol. 2, no. 2, p. 158, Oct. 2016, doi: 10.2196/publichealth.5869.

[15] R. E. Walker, C. R. Keane, and J. G. Burke, "Disparities and Access to Healthy Food in the United States: A Review of Food Deserts Literature," *Health & Place*, vol. 16, no. 5, pp. 876–884, Sep. 2010, doi: 10.1016/j.healthplace.2010.04.013.

[16] D. A. Freedman, "Local Food Environments: They're All Stocked Differently," *American Journal of Community Psychology*, vol. 44, no. 3–4, pp. 382–393, 2009, doi: 10.1007/s10464-009-9272-6.

[17] D. A. Freedman and B. A. Bell, "Access to Healthful Foods among an Urban Food Insecure Population: Perceptions versus Reality," *Journal of Urban Health*, vol. 86, no. 6, pp. 825–838, Nov. 2009, doi: 10.1007/s11524-009-9408-x.

[18] L. B. Lewis *et al.*, “African Americans’ Access to Healthy Food Options in South Los Angeles Restaurants,” *American Journal of Public Health*, vol. 95, no. 4, pp. 668–673, Apr. 2005, doi: 10.2105/AJPH.2004.050260.
[19] K. Morland and S. Filomena, “Disparities in the Availability of Fruits and Vegetables between Racially Segregated Urban Neighbourhoods,” *Public Health Nutrition*, vol. 10, no. 12, pp. 1481–89, Dec. 2007, doi: 10.1017/S1368980007000079.
[20] K. Glanz, J. F. Sallis, B. E. Saelens, and L. D. Frank, “Nutrition Environment Measures Survey in Stores (NEMS-S)Development and Evaluation,” *American Journal of Preventive Medicine*, vol. 32, no. 4, pp. 282–289, Apr. 2007, doi: 10.1016/j.amepre.2006.12.019.
[21] J. P. Block, R. A. Scribner, and K. B. DeSalvo, “Fast Food, Race/Ethnicity, and Income,” *Am J Prev Med*, p. 7, 2004.
[22] S. Inagami, D. A. Cohen, B. K. Finch, and S. M. Asch, “Grocery Store Locations, Weight, and Neighborhoods,” *Am J Prev Med*, p. 8, 2006.
[23] Q. C. Nguyen *et al.*, “Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity,” *Appl Geogr*, vol. 73, pp. 77–88, Aug. 2016, doi: 10.1016/j.apgeog.2016.06.003.
[24] S. S. Sharma and M. D. Choudhury, “Measuring and Characterizing Nutritional Information of Food and Ingestion Content in Instagram,” in *Proceedings of the 24th International Conference on World Wide Web - WWW ’15 Companion*, Florence, Italy, 2015, pp. 115–116. doi: 10.1145/2740908.2742754.
[25] R. West, R. W. White, and E. Horvitz, “From Cookies to Cooks: Insights on Dietary Patterns via Analysis of Web Usage Logs,” *arXiv:1304.3742 [physics]*, Apr. 2013, Accessed: Jan. 21, 2022. [Online]. Available: http://arxiv.org/abs/1304.3742
[26] S. Abbar, Y. Mejova, and I. Weber, *You Tweet What You Eat: Studying Food Consumption Through Twitter*. 2015. doi: 10.1145/2702123.2702153.
[27] J. K. Ward *et al.*, “The French public’s attitudes to a future COVID-19 vaccine: The politicization of a public health issue,” *Social Science & Medicine*, vol. 265, p. 113414, Nov. 2020, doi: 10.1016/j.socscimed.2020.113414.
[28] K. Wang *et al.*, “Intention of nurses to accept coronavirus disease 2019 vaccination and change of intention to accept seasonal influenza vaccination during the coronavirus disease 2019 pandemic: A cross-sectional survey,” *Vaccine*, vol. 38, no. 45, pp. 7049–7056, Oct. 2020, doi: 10.1016/j.vaccine.2020.09.021.
[29] J. B. Ruiz and R. A. Bell, “Predictors of intention to vaccinate against COVID-19: Results of a nationwide survey,” *Vaccine*, vol. 39, no. 7, pp. 1080–1086, Feb. 2021, doi: 10.1016/j.vaccine.2021.01.010.
[30] P. L. Reiter, M. L. Pennell, and M. L. Katz, “Acceptability of a COVID-19 vaccine among adults in the United States: How many people would get vaccinated?,” *Vaccine*, vol. 38, no. 42, pp. 6500–6507, Sep. 2020, doi: 10.1016/j.vaccine.2020.08.043.
[31] S. Bell, R. Clarke, S. Mounier-Jack, J. L. Walker, and P. Paterson, “Parents’ and guardians’ views on the acceptability of a future COVID-19 vaccine: A multi-methods study in England,” *Vaccine*, vol. 38, no. 49, pp. 7789–7798, Nov. 2020, doi: 10.1016/j.vaccine.2020.10.027.
[32] N. S. Sattar and S. Arifuzzaman, “COVID-19 Vaccination Awareness and Aftermath: Public Sentiment Analysis on Twitter Data and Vaccinated Population Prediction in the USA,” *Applied Sciences*, vol. 11, no. 13, Art. no. 13, Jan. 2021, doi: 10.3390/app11136128.
[33] S. Liu and J. Liu, “Public attitudes toward COVID-19 vaccines on English-language Twitter: A sentiment analysis,” *Vaccine*, vol. 39, no. 39, pp. 5499–5505, Sep. 2021, doi: 10.1016/j.vaccine.2021.08.058.
[34] A. Hussain *et al.*, “Artificial Intelligence–Enabled Analysis of Public Attitudes on Facebook and Twitter Toward COVID-19 Vaccines in the United Kingdom and the United States:

Observational Study," *Journal of Medical Internet Research*, vol. 23, no. 4, p. e26627, Apr. 2021, doi: 10.2196/26627.
[35] S. Yousefinaghani, R. Dara, S. Mubareka, A. Papadopoulos, and S. Sharif, "An analysis of COVID-19 vaccine sentiments and opinions on Twitter," *International Journal of Infectious Diseases*, vol. 108, pp. 256–262, Jul. 2021, doi: 10.1016/j.ijid.2021.05.059.
[36] M. T. J. Ansari and N. A. Khan, "Worldwide COVID-19 Vaccines Sentiment Analysis Through Twitter Content," *ELECTRON J GEN MED*, vol. 18, no. 6, p. em329, Nov. 2021, doi: 10.29333/ejgm/11316.
[37] J. C. Lyu, E. L. Han, and G. K. Luli, "COVID-19 Vaccine–Related Discussion on Twitter: Topic Modeling and Sentiment Analysis," *Journal of Medical Internet Research*, vol. 23, no. 6, p. e24435, Jun. 2021, doi: 10.2196/24435.
[38] L. C. Jiang, T. H. Chu, and M. Sun, "Characterization of Vaccine Tweets During the Early Stage of the COVID-19 Outbreak in the United States: Topic Modeling Analysis," *JMIR Infodemiology*, vol. 1, no. 1, p. e25636, Sep. 2021, doi: 10.2196/25636.
[39] S. Yousefinaghani, R. Dara, S. Mubareka, and S. Sharif, "Prediction of COVID-19 Waves Using Social Media and Google Search: A Case Study of the US and Canada," *Front Public Health*, vol. 9, p. 656635, Apr. 2021, doi: 10.3389/fpubh.2021.656635.
[40] D. E. O'Leary and V. C. Storey, "A Google–Wikipedia–Twitter Model as a Leading Indicator of the Numbers of Coronavirus Deaths," *Intelligent Systems in Accounting, Finance and Management*, vol. 27, no. 3, pp. 151–158, 2020, doi: 10.1002/isaf.1482.
[41] J. Sun and P. A. Gloor, "Assessing the Predictive Power of Online Social Media to Analyze COVID-19 Outbreaks in the 50 U.S. States," *Future Internet*, vol. 13, no. 7, Art. no. 7, Jul. 2021, doi: 10.3390/fi13070184.
[42] C. Shen, A. Chen, C. Luo, J. Zhang, B. Feng, and W. Liao, "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, May 2020, doi: 10.2196/19421.
[43] P. Cihan, "Forecasting fully vaccinated people against COVID-19 and examining future vaccination rate for herd immunity in the US, Asia, Europe, Africa, South America, and the World," *Appl Soft Comput*, vol. 111, p. 107708, Nov. 2021, doi: 10.1016/j.asoc.2021.107708.
[44] Q. Cheong, M. Au-yeung, S. Quon, K. Concepcion, and J. D. Kong, "Predictive Modeling of Vaccination Uptake in US Counties: A Machine Learning–Based Approach," *Journal of Medical Internet Research*, vol. 23, no. 11, p. e33231, Nov. 2021, doi: 10.2196/33231.
[45] A. Shaham, G. Chodick, V. Shalev, and D. Yamin, "Personal and social patterns predict influenza vaccination decision," *BMC Public Health*, vol. 20, no. 1, p. 222, Feb. 2020, doi: 10.1186/s12889-020-8327-3.
[46] N. Mannion, "Predictions of Change in Child Immunization Rates Using an Automated Approach: USA," p. 28.
[47] "Enabling the future of academic research with the Twitter API." https://blog.twitter.com/developer/en_us/topics/tools/2021/enabling-the-future-of-academic-research-with-the-twitter-api (accessed Feb. 07, 2022).
[48] Y. Kim, R. Nordgren, and S. Emery, "The Story of Goldilocks and Three Twitter's APIs: A Pilot Study on Twitter Data Sources and Disclosure," *Int J Environ Res Public Health*, vol. 17, no. 3, p. 864, Feb. 2020, doi: 10.3390/ijerph17030864.
[49] C. Gerlitz and B. Rieder, "Mining One Percent of Twitter: Collections, Baselines, Sampling," *M/C Journal*, vol. 16, no. 2, Art. no. 2, Mar. 2013, doi: 10.5204/mcj.620.
[50] "API reference index." https://developer.twitter.com/en/docs/api-reference-index (accessed Feb. 07, 2022).
[51] "Enterprise." https://developer.twitter.com/en/docs/twitter-api/enterprise (accessed Feb. 07, 2022).

[52] “Twitter API v2 tools & libraries.” https://developer.twitter.com/en/docs/twitter-api/tools-and-libraries/v2 (accessed Feb. 17, 2022).
[53] “Filtering Tweets by location.” https://developer.twitter.com/en/docs/tutorials/filtering-tweets-by-location (accessed Feb. 17, 2022).
[54] “Using standard search.” https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/guides/standard-operators (accessed Feb. 17, 2022).
[55] “Tweet object.” https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet (accessed Feb. 17, 2022).
[56] D. M. E.-D. M. Hussein, “A survey on sentiment analysis challenges,” *Journal of King Saud University - Engineering Sciences*, vol. 30, no. 4, pp. 330–338, Oct. 2018, doi: 10.1016/j.jksues.2016.04.002.
[57] M. L. Loureiro and M. Alló, “Sensing climate change and energy issues: Sentiment and emotion analysis with social media in the U.K. and Spain,” *Energy Policy*, vol. 143, p. 111490, Aug. 2020, doi: 10.1016/j.enpol.2020.111490.
[58] M. Naldi, “A review of sentiment computation methods with R packages,” *arXiv:1901.08319 [cs]*, Jan. 2019, Accessed: Jan. 13, 2022. [Online]. Available: http://arxiv.org/abs/1901.08319
[59] E. Gabarron, E. Dorronzoro, O. Rivera-Romero, and R. Wynn, “Diabetes on Twitter: A Sentiment Analysis,” *J Diabetes Sci Technol*, vol. 13, no. 3, pp. 439–444, May 2019, doi: 10.1177/1932296818811679.
[60] X. Ji, S. A. Chun, Z. Wei, and J. Geller, “Twitter sentiment classification for measuring public health concerns,” *Soc. Netw. Anal. Min.*, vol. 5, no. 1, p. 13, Dec. 2015, doi: 10.1007/s13278-015-0253-5.
[61] F. Greaves, D. Ramirez-Cano, C. Millett, A. Darzi, and L. Donaldson, “Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online,” *Journal of Medical Internet Research*, vol. 15, no. 11, p. e2721, Nov. 2013, doi: 10.2196/jmir.2721.
[62] J. Azétsop and T. R. Joy, “Access to nutritious food, socioeconomic individualism and public health ethics in the USA: a common good approach,” *Philos Ethics Humanit Med*, vol. 8, p. 16, Oct. 2013, doi: 10.1186/1747-5341-8-16.
[63] *USDA ERS - Definitions of Food Security, 2019*. [Online]. Available: https://www.ers.usda.gov/topics/food-nutrition-assistance/food-security-in-the-us/definitions-of-food-security/.
[64] D. Choudhury, S. S. Munmun, and E. Kiciman, “Characterizing Dietary Choices, Nutrition, and Language in Food Deserts via Social Media,” in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW*, San Francisco, California, USA, 2016, vol. 16, pp. 1155–1168. doi: 10.1145/2818048.2819956.
[65] J. Crowe, C. Lacy, and Y. Columbus, “Barriers to Food Security and Community Stress in an Urban Food Desert,” *Urban Science*, vol. 2, no. 2, Art. no. 2, Jun. 2018, doi: 10.3390/urbansci2020046.
[66] G. Sansom and B. Hannibal, “Disparate access to nutritional food; place, race and equity in the United States,” *BMC Nutrition*, vol. 7, no. 1, p. 29, Jun. 2021, doi: 10.1186/s40795-021-00434-2.
[67] A. Murrell and R. Jones, “Measuring Food Insecurity Using the Food Abundance Index: Implications for Economic, Health and Social Well-Being,” *Int J Environ Res Public Health*, vol. 17, no. 7, p. 2434, Apr. 2020, doi: 10.3390/ijerph17072434.
[68] V. G. V. Vydiswaran *et al.*, “Uncovering the Relationship between Food-Related Discussion on Twitter and Neighborhood Characteristics,” *Journal of the American Medical Informatics Association*, doi: 10.1093/jamia/ocz181.

[69] R. J. Gore, S. Diallo, and J. Padilla, “You Are What You Tweet: Connecting the Geographic Variation in America’s Obesity Rate to Twitter Content,” *PLOS ONE*, vol. 10, no. 9, p. e0133505, Sep. 2015, doi: 10.1371/journal.pone.0133505.
[70] U. C. Bureau, “City and Town Population Totals: 2010-2019,” *Census.gov*. https://www.census.gov/data/tables/time-series/demo/popest/2010s-total-cities-and-towns.html (accessed Jan. 29, 2022).
[71] C. on P. H. S. to I. Health and I. of Medicine, *Funding Sources and Structures to Build Public Health*. National Academies Press (US), 2012. Accessed: Oct. 18, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK201025/
[72] “FoodData Central.” https://fdc.nal.usda.gov/ (accessed Apr. 19, 2022).
[73] “List of fast food restaurant chains,” *Wikipedia*. Apr. 12, 2022. Accessed: Apr. 24, 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=List_of_fast_food_restaurant_chains&oldid=1082336902
[74] J. Silge and D. Robinson, “tidytext: Text Mining and Analysis Using Tidy Data Principles in R,” *JOSS*, vol. 1, no. 3, p. 37, Jul. 2016, doi: 10.21105/joss.00037.
[75] Z. Zhang, D. Yin, K. Virrantaus, X. Ye, and S. Wang, “Modeling human activity dynamics: an object-class oriented space–time composite model based on social media and urban infrastructure data,” *Comput.Urban Sci.*, vol. 1, no. 1, p. 7, May 2021, doi: 10.1007/s43762-021-00006-x.
[76] *USDA ERS - Documentation*. [Online]. Available: https://www.ers.usda.gov/data-products/food-access-research-atlas/documentation/#lowaccess.
[77] P. Dutko, M. V. Ploeg, and T. Farrigan, *Characteristics and Influential Factors of Food Deserts*, vol. n.d., 36.
[78] X. H. Cao, I. Stojkovic, and Z. Obradovic, “A robust data scaling algorithm to improve classification accuracies in biomedical data,” *BMC Bioinformatics*, vol. 17, no. 1, Sep. 2016, doi: 10.1186/s12859-016-1236-x.
[79] M. Kuhn, *The caret Package*. Accessed: Apr. 19, 2022. [Online]. Available: https://topepo.github.io/caret/
[80] C. Halimu, A. Kasem, and S. H. S. Newaz, “Empirical Comparison of Area under ROC curve (AUC) and Mathew Correlation Coefficient (MCC) for Evaluating Machine Learning Algorithms on Imbalanced Datasets for Binary Classification,” in *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing*, New York, NY, USA, Jan. 2019, pp. 1–6. doi: 10.1145/3310986.3311023.
[81] L. Mitchell, M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth, “The geography of happiness: connecting twitter sentiment and expression, demographics, and objective characteristics of place,” *PLoS One*, vol. 8, no. 5, p. e64417, 2013, doi: 10.1371/journal.pone.0064417.
[82] J. Helliwell, R. Layard, J. Sacks, J.-E. Neve, L. Atkin, and S. Wang, “World Happiness Report 2022,” *Happiness and Subjective Well-Being*, Jan. 2022, [Online]. Available: https://www.wellbeingintlstudiesrepository.org/hw_happiness/2
[83] P. Xu, M. Dredze, and D. A. Broniatowski, “The Twitter Social Mobility Index: Measuring Social Distancing Practices With Geolocated Tweets,” *Journal of Medical Internet Research*, vol. 22, no. 12, p. e21499, Dec. 2020, doi: 10.2196/21499.
[84] 1615 L. St NW, Suite 800Washington, and D. 20036USA202-419-4300 | M.-857-8562 | F.-419-4372 | M. Inquiries, “Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018,” *Pew Research Center*. https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/ (accessed Apr. 15, 2021).

[85] U. C. Bureau, “Household Pulse Survey COVID-19 Vaccination Tracker,” *Census.gov*. https://www.census.gov/library/visualizations/interactive/household-pulse-survey-covid-19-vaccination-tracker.html (accessed Jan. 13, 2022).
[86] R. Garett and S. D. Young, “Online misinformation and vaccine hesitancy,” *Transl Behav Med*, vol. 11, no. 12, pp. 2194–2199, Dec. 2021, doi: 10.1093/tbm/ibab128.
[87] Y. Pershad, P. T. Hangge, H. Albadawi, and R. Oklu, “Social Medicine: Twitter in Healthcare,” *Journal of Clinical Medicine*, vol. 7, no. 6, Art. no. 6, Jun. 2018, doi: 10.3390/jcm7060121.
[88] R. Chunara, J. R. Andrews, and J. S. Brownstein, “Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak,” *Am J Trop Med Hyg*, vol. 86, no. 1, pp. 39–45, Jan. 2012, doi: 10.4269/ajtmh.2012.11-0597.
[89] S. J. Ball-Rokeach and M. L. DeFleur, “A Dependency Model of Mass-Media Effects,” *Communication Research*, vol. 3, no. 1, pp. 3–21, Jan. 1976, doi: 10.1177/009365027600300101.
[90] M. Charquero-Ballester, J. G. Walter, I. A. Nissen, and A. Bechmann, “Different types of COVID-19 misinformation have different emotional valence on Twitter,” *Big Data & Society*, vol. 8, no. 2, p. 20539517211041280, Jul. 2021, doi: 10.1177/20539517211041279.
[91] E. Bonnevie, A. Gallegos-Jeffrey, J. Goldbarg, B. Byrd, and J. Smyser, “Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic,” *Journal of Communication in Healthcare*, vol. 14, no. 1, pp. 12–19, Jan. 2021, doi: 10.1080/17538068.2020.1858222.
[92] M. S. Deiner *et al.*, “Facebook and Twitter vaccine sentiment in response to measles outbreaks,” *Health Informatics J*, vol. 25, no. 3, pp. 1116–1132, Sep. 2019, doi: 10.1177/1460458217740723.
[93] D. Arigo, S. Pagoto, L. Carter-Harris, S. E. Lillie, and C. Nebeker, “Using social media for health research: Methodological and ethical considerations for recruitment and intervention delivery,” *DIGITAL HEALTH*, vol. 4, p. 2055207618771757, Jan. 2018, doi: 10.1177/2055207618771757.
[94] K. B. Wright, “Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services,” *Journal of Computer-Mediated Communication*, vol. 10, no. 3, p. JCMC1034, Apr. 2005, doi: 10.1111/j.1083-6101.2005.tb00259.x.
[95] D. Enriquez and A. Goldstein, “COVID-19’s Socioeconomic Impact on Low-Income Benefit Recipients: Early Evidence from Tracking Surveys,” *Socius*, vol. 6, p. 2378023120970794, Jan. 2020, doi: 10.1177/2378023120970794.
[96] D. X. Morales, S. A. Morales, and T. F. Beltran, “Food Insecurity in Households with Children Amid the COVID-19 Pandemic: Evidence from the Household Pulse Survey,” *Social Currents*, vol. 8, no. 4, pp. 314–325, Aug. 2021, doi: 10.1177/23294965211011593.
[97] M. Daly, A. Jones, and E. Robinson, “Public Trust and Willingness to Vaccinate Against COVID-19 in the US From October 14, 2020, to March 29, 2021,” *JAMA*, vol. 325, no. 23, pp. 2397–2399, Jun. 2021, doi: 10.1001/jama.2021.8246.
[98] “NRC Emotion Lexicon.” http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm (accessed Jan. 13, 2022).
[99] M. Abu Kausar, A. Soosaimanickam, and M. Nasar, “Public Sentiment Analysis on Twitter Data during COVID-19 Outbreak,” *International Journal of Advanced Computer Science and Applications*, vol. 12, Jan. 2021, doi: 10.14569/IJACSA.2021.0120252.
[100] M. D. at J. H. University, “Data.” http://twitterdata.covid19dataresources.org/data/ (accessed Dec. 04, 2022).

[101]W. Haynes, “Holm’s Method,” in *Encyclopedia of Systems Biology*, W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, Eds. New York, NY: Springer, 2013, pp. 902–902. doi: 10.1007/978-1-4419-9863-7_1214.
[102]A. Bari *et al.*, “Exploring Coronavirus Disease 2019 Vaccine Hesitancy on Twitter Using Sentiment Analysis and Natural Language Processing Algorithms,” *Clin Infect Dis*, vol. 74, no. Suppl_3, pp. e4–e9, May 2022, doi: 10.1093/cid/ciac141.
[103]T. W. House, “Remarks by President Biden Marking the 150 Millionth COVID-19 Vaccine Shot,” *The White House*, Apr. 06, 2021. https://www.whitehouse.gov/briefing-room/speeches-remarks/2021/04/06/remarks-by-president-biden-marking-the-150-millionth-covid-19-vaccine-shot/ (accessed Aug. 07, 2022).
[104]J. Huang, S. Kumar, and C. Hu, “A Literature Review of Online Identity Reconstruction,” *Frontiers in Psychology*, vol. 12, 2021, Accessed: Aug. 10, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.696552
[105]H.-W. Kim, J. R. Zheng, and S. Gupta, “Examining knowledge contribution from the perspective of an online identity in blogging communities,” *Computers in Human Behavior*, vol. 27, no. 5, pp. 1760–1770, Sep. 2011, doi: 10.1016/j.chb.2011.03.003.
[106]C. Hu, L. Zhao, and J. Huang, “Achieving self-congruency? Examining why individuals reconstruct their virtual identity in communities of interest established within social network platforms,” *Computers in Human Behavior*, vol. 50, pp. 465–475, Sep. 2015, doi: 10.1016/j.chb.2015.04.027.
[107]J. A. Bargh, K. Y. A. McKenna, and G. M. Fitzsimons, “Can You See the Real Me? Activation and Expression of the ‘True Self’ on the Internet,” *Journal of Social Issues*, vol. 58, no. 1, pp. 33–48, 2002, doi: 10.1111/1540-4560.00247.
[108]J. S. Donath, “Identity and deception in the virtual community,” in *Communities in Cyberspace*, Routledge, 1998.
[109]H. Jensen Schau and M. C. Gilly, “We Are What We Post? Self-Presentation in Personal Web Space,” *Journal of Consumer Research*, vol. 30, no. 3, pp. 385–404, Dec. 2003, doi: 10.1086/378616.
[110]V. Govindan and V. Balakrishnan, “A machine learning approach in analysing the effect of hyperboles using negative sentiment tweets for sarcasm detection,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, Part A, pp. 5110–5120, Sep. 2022, doi: 10.1016/j.jksuci.2022.01.008.
[111]T. Koyama, D. Weeraratne, J. L. Snowdon, and L. Parida, “Emergence of Drift Variants That May Affect COVID-19 Vaccine Development and Antibody Treatment,” *Pathogens*, vol. 9, no. 5, Art. no. 5, May 2020, doi: 10.3390/pathogens9050324.
[112]D. M. Morens, G. K. Folkers, and A. S. Fauci, “The Concept of Classical Herd Immunity May Not Apply to COVID-19,” *The Journal of Infectious Diseases*, vol. 226, no. 2, pp. 195–198, Jul. 2022, doi: 10.1093/infdis/jiac109.
[113]“NIH experts discuss controlling COVID-19 in commentary on herd immunity,” *National Institutes of Health (NIH)*, Mar. 31, 2022. https://www.nih.gov/news-events/news-releases/nih-experts-discuss-controlling-covid-19-commentary-herd-immunity (accessed Sep. 12, 2022).
[114]A. Bhattacharjee, G. K. Vishwakarma, N. Gajare, and N. Singh, “Time Series Analysis Using Different Forecast Methods and Case Fatality Rate for Covid-19 Pandemic,” *Regional Science Policy & Practice*, vol. n/a, no. n/a, doi: 10.1111/rsp3.12555.
[115]M. Kim, “Prediction of COVID-19 Confirmed Cases after Vaccination: Based on Statistical and Deep Learning Models,” *SciMedicine Journal*, vol. 3, no. 2, Art. no. 2, Jun. 2021, doi: 10.28991/SciMedJ-2021-0302-7.
[116]M. R. Davahli, W. Karwowski, and K. Fiok, “Optimizing COVID-19 vaccine distribution across the United States using deterministic and stochastic recurrent neural networks,” *PLOS ONE*, vol. 16, no. 7, p. e0253925, Jul. 2021, doi: 10.1371/journal.pone.0253925.

[117]F. Germani and N. Biller-Andorno, “The anti-vaccination infodemic on social media: A behavioral analysis,” *PLOS ONE*, vol. 16, no. 3, p. e0247642, Mar. 2021, doi: 10.1371/journal.pone.0247642.
[118]A. L. Schmidt, F. Zollo, A. Scala, C. Betsch, and W. Quattrociocchi, “Polarization of the vaccination debate on Facebook,” *Vaccine*, vol. 36, no. 25, pp. 3606–3612, Jun. 2018, doi: 10.1016/j.vaccine.2018.05.040.
[119]N. Smith and T. Graham, “Mapping the anti-vaccination movement on Facebook,” *Information, Communication & Society*, vol. 22, no. 9, pp. 1310–1327, Jul. 2019, doi: 10.1080/1369118X.2017.1418406.
[120]U. C. Bureau, “Metropolitan and Micropolitan Statistical Areas Population Totals and Components of Change: 2020-2021,” *Census.gov*. https://www.census.gov/data/tables/time-series/demo/popest/2020s-total-metro-and-micro-statistical-areas.html (accessed Aug. 07, 2022).
[121]S. Melotte and M. Kejriwal, “Predicting zip code-level vaccine hesitancy in US Metropolitan Areas using machine learning models on public tweets,” *PLOS Digital Health*, vol. 1, no. 4, p. e0000021, Apr. 2022, doi: 10.1371/journal.pdig.0000021.
[122]M. Malik, H. Lamba, C. Nakos, and J. Pfeffer, “Population Bias in Geotagged Tweets,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 9, no. 4, Art. no. 4, 2015.
[123]“COVID-19 Vaccinations in the United States,County | Data | Centers for Disease Control and Prevention.” https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh (accessed Jan. 13, 2022).
[124]U. C. Bureau, “Delineation Files,” *Census.gov*. https://www.census.gov/geographies/reference-files/time-series/demo/metro-micro/delineation-files.html (accessed Aug. 07, 2022).
[125]Z. Liu, L. Liu, and H. Li, “Determinants of information retweeting in microblogging,” *Internet Research*, vol. 22, no. 4, pp. 443–466, Jan. 2012, doi: 10.1108/10662241211250980.
[126]S. Gittelman *et al.*, “A New Source of Data for Public Health Surveillance: Facebook Likes,” *Journal of Medical Internet Research*, vol. 17, no. 4, p. 98, 2015, doi: 10.2196/jmir.3970.
[127]A. Talaei-Khoei, J. M. Wilson, and S.-F. Kazemi, “Period of Measurement in Time-Series Predictions of Disease Counts from 2007 to 2017 in Northern Nevada: Analytics Experiment,” *JMIR Public Health and Surveillance*, vol. 5, no. 1, p. e11357, Jan. 2019, doi: 10.2196/11357.
[128]S. I. Alzahrani, I. A. Aljamaan, and E. A. Al-Fakih, “Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions,” *J Infect Public Health*, vol. 13, no. 7, pp. 914–919, Jul. 2020, doi: 10.1016/j.jiph.2020.06.001.
[129]A. Zheng, Q. Fang, Y. Zhu, C. Jiang, F. Jin, and X. Wang, “An application of ARIMA model for predicting total health expenditure in China from 1978-2022,” *J Glob Health*, vol. 10, no. 1, p. 010803, doi: 10.7189/jogh.10.010803.
[130]A. Maugeri, M. Barchitta, and A. Agodi, “Using Google Trends to Predict COVID-19 Vaccinations and Monitor Search Behaviours about Vaccines: A Retrospective Analysis of Italian Data,” *Vaccines (Basel)*, vol. 10, no. 1, p. 119, Jan. 2022, doi: 10.3390/vaccines10010119.
[131]R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications: With R Examples*. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-52452-8.
[132]A. T. Jebb, L. Tay, W. Wang, and Q. Huang, “Time series analysis for psychological research: examining and forecasting change,” *Front Psychol*, vol. 6, p. 727, Jun. 2015, doi: 10.3389/fpsyg.2015.00727.

[133]"forecast package - RDocumentation." https://www.rdocumentation.org/packages/forecast/versions/8.16 (accessed Aug. 07, 2022).
[134]R. J. Hyndman and Y. Khandakar, "Automatic Time Series Forecasting: The forecast Package for R," *Journal of Statistical Software*, vol. 27, pp. 1–22, Jul. 2008, doi: 10.18637/jss.v027.i03.
[135]J. Zhang, R. Shang, C. Rittenhouse, C. Witharana, and Z. Zhu, "Evaluating the impacts of models, data density and irregularity on reconstructing and forecasting dense Landsat time series," *Science of Remote Sensing*, vol. 4, p. 100023, Dec. 2021, doi: 10.1016/j.srs.2021.100023.
[136]B. Jiang, S. Liang, J. Wang, and Z. Xiao, "Modeling MODIS LAI time series using three statistical methods," *Remote Sensing of Environment*, vol. 114, no. 7, pp. 1432–1444, Jul. 2010, doi: 10.1016/j.rse.2010.01.026.
[137]H. Alabdulrazzaq, M. N. Alenezi, Y. Rawajfih, B. A. Alghannam, A. A. Al-Hassan, and F. S. Al-Anzi, "On the accuracy of ARIMA based prediction of COVID-19 spread," *Results in Physics*, vol. 27, p. 104509, Aug. 2021, doi: 10.1016/j.rinp.2021.104509.
[138]R. Alegado and G. Tumibay, "Forecasting Measles Immunization Coverage Using ARIMA Model," *Journal of Computer and Communications*, vol. 07, pp. 157–168, Jan. 2019, doi: 10.4236/jcc.2019.710015.
[139]"Full article: Forecasting the COVID-19 vaccine uptake rate: an infodemiological study in the US." https://www.tandfonline.com/doi/full/10.1080/21645515.2021.2017216 (accessed Aug. 07, 2022).
[140]J. Y. Breland, L. M. Quintiliani, K. L. Schneider, C. N. May, and S. Pagoto, "Social Media as a Tool to Increase the Impact of Public Health Research," *Am J Public Health*, vol. 107, no. 12, pp. 1890–1891, Dec. 2017, doi: 10.2105/AJPH.2017.304098.
[141]E. M. Goldberg *et al.*, "Using Social Media for Clinical Research: Recommendations and Examples From the Brown-Lifespan Center for Digital Health," *Journal of Medical Internet Research*, vol. 24, no. 6, p. e35804, Jun. 2022, doi: 10.2196/35804.
[142]D. A. Ogundijo, A. A. Tas, and B. A. Onarinde, "Age, an Important Sociodemographic Determinant of Factors Influencing Consumers' Food Choices and Purchasing Habits: An English University Setting," *Frontiers in Nutrition*, vol. 9, 2022, Accessed: Nov. 30, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnut.2022.858593
[143]E. Di Minin, C. Fink, A. Hausmann, J. Kremer, and R. Kulkarni, "How to address data privacy concerns when using social media data in conservation science," *Conservation Biology*, vol. 35, no. 2, pp. 437–446, 2021, doi: 10.1111/cobi.13708.
[144]M. Zimmer, "'But the data is already public': on the ethics of research in Facebook," in *The Ethics of Information Technologies*, Routledge, 2017.