THE ANALYSIS OF USER CHARACTERISTICS ON TWITTER DURING EARLY STAGE OF THE COVID-19 PANDEMIC: A COMPARISON STUDY BEFORE AND AFTER DECLARATION OF THE COVID-19 PANDEMIC

by

Mutasim Abdulrahman Alfadhel

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Information Studies

at

The University of Wisconsin-Milwaukee

May 2022

ABSTRACT

THE ANALYSIS OF USER CHARACTERISTICS ON TWITTER DURING EARLY STAGE OF THE COVID-19 PANDEMIC: A COMPARISON STUDY BEFORE AND AFTER DECLARATION OF THE COVID-19 PANDEMIC

by

Mutasim Abdulrahman Alfadhel

The University of Wisconsin-Milwaukee, 2022
Under the Supervision of Professor Xiangming "Simon" Mu

In December 2019, the coronavirus disease 2019 (Covid-19) was officially reported as an acute respiratory infection, which was first identified in Wuhan, China. On March 11th, 2020, the World Health Organization (WHO) declared that Covid-19 could be characterized as a pandemic. Governments across the world imposed or recommended various non-medical interventions to reduce transmission of Covid-19, such as washing hands, wearing face masks, social distancing, and quarantining as well as lockdown measures including banning large gatherings, issuing stay-at-home orders, closing certain businesses, and imposing travel restrictions. The increase in social, behavioral, and economic issues that the disease has generated has had both negative effects, such as large numbers of new cases and mortality at its peak times and locations, and positive effects, such as scientific discoveries. Given the enormous and lasting impact of the pandemic, the primary aim of this study was to explore the information flow, users' perceptions and sentiments and topics dominating social media as captured in Twitter in the 60 days before and the 60 days after the declaration of the Covid-19 pandemic and analyze these data using machine learning techniques.

Over 68 million tweets for this study were collected from Georgia State University's Panacea Lab. Before being analyzed, the data were prepared and cleansed. After cleansing the

data, 21,655,284 tweets were used for data analysis and categorized into tweets posted during the 60 days before the declaration of the Covid-19 pandemic (3,406,055) and tweets posted during the 60 days after the declaration (18,249,229). Three machine learning techniques and inferential analysis were applied. The sentiment analysis, and emotional analysis used to understand users' characteristics, a Latent Dirichlet Allocation (LDA) model was employed to uncover discussion topics about the Covid-19 pandemic that users tweeted. Inferential analysis was applied to investigate the differences between sentiment characteristics and discussion topics.

The results of the sentiment characteristics and emotional analysis show how the declaration of the Covid-19 pandemic change the users' characteristics during the Covid-19 pandemic on Twitter. The results of the LDA model show various discussion topics 60 days before and 60 days after the declaration of the Covid-19 pandemic. The inferential analysis showed there are no differences between the two groups.

The findings of this study revealed the characteristics of users on Twitter in the early stage of the Covid-19 pandemic. Theoretically, the findings demonstrated the significance of social media for policymakers in monitoring a health crisis. The distinctive implication of this study was to identify users' characteristics on Twitter as a source of information and knowledge relevant to managing a health crisis for policymakers. The methodology employed in this study demonstrated a systematic approach to evaluating Twitter users' behaviors in the early stage of the Covid-19 pandemic. The outcomes of this study can contribute to establishing a fast and low-human-effort surveillance system for monitoring people's attitudes towards policies and policymakers during a health or other crisis.

DEDICATION

To,

My parents,

My wife,

My sons

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to praise and thank God “Allah (SWT) ,” the almighty, for giving me countless blessings, knowledge, strength, and courage to accomplish this dissertation study.

Then I would like to express my deep and sincere gratitude to my advisor, Dr. Xiangming "Simon" Mu, and thank him for his patience, mentorship, guidance, encouragement, and endless support during my Ph.D. journey. The skills and knowledge I learned from him allowed me to master the academic skills that I used during the research and will continue to use in the future. I would like to express my deepest appreciation to my committee members. I am highly thankful to Prof. Jin Zhang for his unwavering support and remarks. I am also very grateful to Dr. Wonchan Choi for his valuable guidance and comments. I am also very thankful to Dr. Min Wu for his suggestions and feedback.

Last but not least, I would like to thank my mother, Fouziah Alfawzan, and my father, Abdulrahman Alfadhel, for their kindness and generosity, for their support throughout my studies, and for their guidance and advice in my life. I would also like to thank my brother and sisters for all of their support throughout my years in the U.S.; my admiration and love for them know no bounds. Moreover, I would like to thank my wife, Sara Alhomaidhi, for her tremendous and endless support and love during my studies. Finally, I am grateful for my children, Rakan and Abdulrahaman, who have been an endless source of joy during the writing of this dissertation.

Chapter 1. Introduction:

Background:

In December 2019, the coronavirus disease 2019 (COVID-19) was first officially reported as an acute respiratory infection, initially identified in Wuhan, China ("Centers for Disease Control and Prevention," 2020). On March 11th, 2020, the World Health Organization (WHO) declared that Covid-19 could be characterized as a pandemic (WHO, 2020a). According to World Health Organization (WHO, 2021b), as of March 16th, 2021, more than 119 million people worldwide had been infected by Covid-19, and more than 2.6 million people had died. Across the world Governments imposed or urged various non-medical actions to reduce transmission of Covid-19, such as washing hands, wearing face masks, social distancing, and quarantining, as well as such lockdown measures as banning gatherings, issuing stay-at-home orders, closing some businesses, and imposing travel restrictions (Gostin & Wiley, 2020). On December 11, 2020, the U.S. Food and Drug Administration (FDA) issued the first emergency use authorization (EUA) for a vaccine for the prevention of COVID-19 in individuals 16 years of age and older, allowing the Pfizer-BioNTech COVID-19 Vaccine to be distributed in the U.S (FDA, 2020).The problems surrounding the pandemic have made its effects increasingly challenging and complicated.

There are various complexities that come not from the Covid-19 itself, but it is from other factors. The increase of the social (Hazel, 2020), behavioral (Knell et al., 2020), and economic (Milani, 2020) issues that the disease has generated and made a negative impact, such as rising numbers of new cases and mortality, and positive impact, such as scientific discoveries, as well as government actions. The reporting of social behaviors also has shown the effects of people's

actions, such as panic buying and food hoarding resulting in shortages, decline of businesses leading to negative economic predictions, and changes in individuals' daily lives such as isolation (Arafat et al., 2020; Gostin & Wiley, 2020). These challenges have many unprecedented aspects and consequences, which scholars in various fields have started to investigate, and there is increasing recognition of the need for multidisciplinary research efforts to fully understand the response to the COVID-19 pandemic. One source of data for these inquiries is the use of social media in emergency situations, particularly when people's direct social interactions are restricted.

People use different social media platforms to share news, information, opinions, and emotions. In previous infectious disease outbreaks , such as Ebola (Odlum & Yoon, 2015) and Zika (Zhang et al., 2019), people used social media platforms in ways similar to their use during Covid-19. Thus, social media provide an essential source of information for scholars seeking to understand users' reactions to and concerns about a particular subject, which in the case of a health crisis can generate information and insights beneficial for private and public health sectors and facilitate designing health strategies and well-informed interventions.

Today's social media are highly popular because they offer privacy protection, immediacy, speed, convenience, anonymity, a wide range of information in a cost effective manner, real time relay options, and the capacity to handle a large number of users simultaneously with easy access (Cotten & Gupta, 2004; Prybutok et al., 2014; Skinner et al., 2003). Moorhead et al. (2013) identify many advantages of using social media, such as increased interactions and sharing, availability of tailored information, accessibility and, in the case of health emergency such as the Covid-19 pandemic, immediate access to health information, peer/social/emotional support, and public health surveillance, which gives social media the

potential to influence health policy. In this dissertation study of these effects of a social media platform, Twitter was the main source of data.

There According to *Statista's* (2019) report, *Number of social media users worldwide 2010-2021, a*t about the start of the pandemic, there were are over 2.82 billion Twitter users around the world and that number that number has been growing every year (*Number of social media users worldwide 2010-2021 | Statista*, 2019) Wikipedia (2019) describes Twitter as an American social networking platform created in March 2006 by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams, launched in July of that year. (2019). Users of the platform are able to post and interact with messages called "tweets," which were initially restricted to 280 characters for all languages except Chinese, Korean, and Japanese (2019). Tweets can include hyperlinks to external resources, such as videos or websites (Grajales Iii et al., 2014). Users can also include "hashtags," in tweets, which is a form of information indexing that enables people to search for tweets that are related to a certain topic (Grajales Iii et al., 2014). Nowadays, Twitter is used for many different purposes, such as news, politics, health, and education. (Chae, 2015) reported that Twitter users were producing 500 million tweets daily, a number that has increased in subsequent years. In sum, Twitter enables users and organizations to promote what they want the public to know in a quick and easy manner.

Significance:

Social media platforms' provision of online interactions among citizens and between the public and government officials has become a hallmark of democracy worldwide. This study can be used in the awareness and conversion stages of a health crisis to share timely information support compliance with the decisions of policymakers. Investigations of the potential of the use of Twitter as a direct communication tool bridge the interests of researchers and policymakers,

showing it to be among the most direct influencers of health legislation (Kapp et al., 2015). Thus, this study provides a lens through which researchers can investigate communities' perceptions, which can help scientists and policy makers develop effective policies during a health crisis.

Twitter can provide communication in real-time with low cost for public health surveillance (Achrekar et al., 2011; Hughes, 2016). Authorities can use Twitter to monitor public response to health issues (Kapp et al., 2015), track and monitor disease outbreaks (Abd-Alrazaq et al., 2020; Kostkova et al., 2010; Liang et al., 2019; Odlum & Yoon, 2015), and identify dissemination of health misinformation (Al-Rakhami & Al-Amri, 2020; Lyu & Luli, 2021). This study provides an opportunity to help accurately monitor public response to health issues, track and monitor disease outbreaks, and identify dissemination of health misinformation. In this study, a comparison of Twitter use during the 60 days before and the 60 days after the declaration using machine learning techniques can provide significant insight into the information flow and users' sentiments and discussion topics in relation to the emergency situation.

Policy makers can take an advantage of Twitter as an effective way to measure the level of healthcare management required during the Covid-19 pandemic. The results of this study may contribute to the improvement of healthcare system informatics by providing information to the healthcare industry and public policy makers that may enable them to clearly see the weaknesses and strengths of decisions made during the Covid-19 pandemic address some shortcomings of current healthcare systems in response to next pandemic. Additionally, the theoretical implications lie in the revealing of sentiment characteristics of users and discussion topics on Twitter during the Covid-19 pandemic. The methodology proposed in this study can be employed to explore users on Twitter focusing on other health outbreaks. In practice, this study

explores the sentiment characteristics and discussion topics derived from tweets related to the Covid-19 pandemic on Twitter. Twitter provides an opportunity for policy makers to track, monitor, and manage the health crisis in effectively and appropriately.

Research Problem, Questions, and Objective:

Social media platforms have become prominent online sources for the exchange of health-related information and advice. Previous studies have dealt mainly with online health information-seeking behaviors, and there have been few investigations of the characteristics of users, the types of information being generated and shared, and users' health related interactions. Specifically, this study is the first comparison of health-related communications on the social platform Twitter 60 days before and 60 days after the first declaration of the Covid-19 pandemic.

Research Problem:

People worldwide were caught unprepared for the Covid-19 pandemic, and the spread of the disease was quickly out of control. This lack of awareness of the novel virus and how to deal with it gave it time to spread. On December 31st, 2019, WHO confirmed the first case of Covid-19 worldwide (United Nations, 2019). On March 9th, 2020, just over two months from that very first case, there were 54,409 confirmed cases and 2,378 deaths globally, and the number had increased tenfold to 583,994 cases by May 4th, 2020 (WHO, 2021). On March 11th, 2020, the WHO officially characterized Covid-19 as a pandemic (WHO, 2020). Covid-19 was not the world's first epidemiologic outbreak, but its rapid spread was fearful. After WHO had characterized Covid-19 as a pandemic, people started to panic, particularly when more and more cases were reported with no travel history or contact with confirmed cases (Bajema et al., 2020). The case of the COVID-19 pandemic shows the substantial effect of this new information environment on social media platforms. The information spreading on social media platforms

can strongly influence users' behavior and alter the effectiveness of the countermeasures deployed by authorities.

After the declaration of a pandemic, many complexities arose, not directly from the infection itself but from related factors, such as increases in social (Hazel, 2020), behavioral (Knell et al., 2020), and financial (FICCI, 2020; "Impact of COVID-19 on Indian Economy: COVID-19 Mitigation Measures Taken by Indian Companies," 2020) issues that the disease generated as numbers of new cases and mortalities climbed, resulting in scientific discoveries and government actions. Thus, people also had different reactions to Covid-19 before and after WHO's declaration that changed the sentiment characteristics of users, and countries employed different measures for dealing with the pandemic. Limited quantitative research on how people's sentiment characteristics and discussion topics towards Covid-19 changes after WHO's declaration of pandemic.

The purpose of this quantitative study was to explore Twitter users' behaviors 60 days before and after the declaration of the pandemic based on Social Cognitive Theory (Bandura, 1986) and using sentiment analysis and topic modeling in order to provide information on the baseline trajectory of people's reactions to Covid19 to help policy makers make appropriate decisions on managing the health crisis and monitoring the level of health consciousness.

Social Cognitive Theory:

Social Cognitive Theory (SCT) is a dynamic, reciprocal model of human behavior in which personal factors, environmental influences, and behavior continually interact (Bandura, 1986). Bandura explains that SCT comprises concepts and processes from cognitive, behavioral, and emotional models of behavior change, so it can be easily implemented in counseling interventions applicable to disease prevention and management. Figure 1 illustrates the three

main factors of SCT. According to Bandura, a basic premise of SCT is that people learn from a combination of their own experiences, observing the actions of others, and then assessing the results of those actions. SCT has four key constructs that are related to health behavior: observational learning, reinforcement, self-control, and self-efficacy (Bandura, 1986).

Bandura implemented SCT to study health promotion and disease prevention. He emphasized that changes in individuals' beliefs in their collective efficacy to achieve social change are a significant component in public health policy strategies for health promotion and disease prevention. In this study, the three components of SCT provide a suitable framework for examining phenomena related to the early stage of the Covid-19 pandemic.

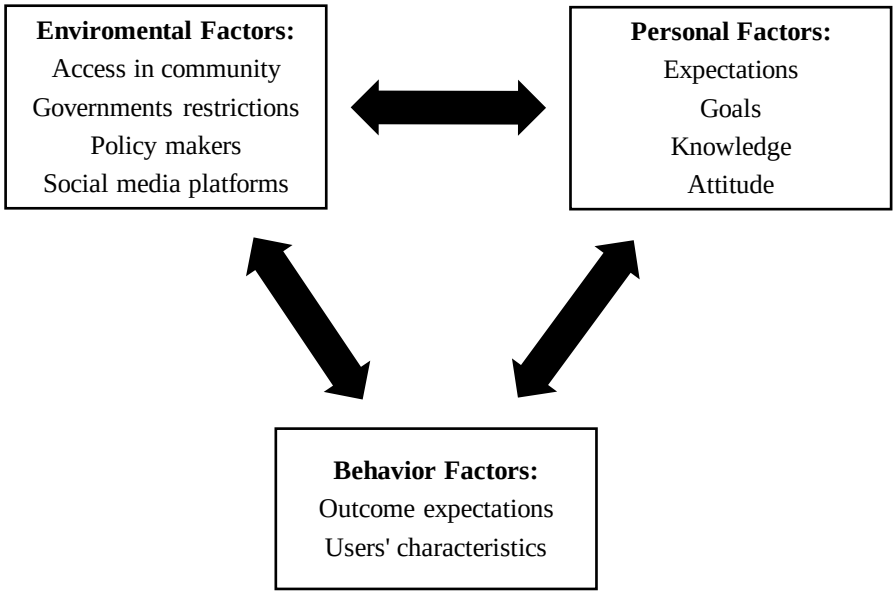

Figure 1. Social Cognitive Theory Factors

Different aspects of environmental factors that influence users on Twitter during the Covid-19 pandemic, such as number of cases, news, and governments restrictions. Users may have different personal factors that can influence their characteristics based on attitude, cultural

background, and level of education. After that, the behavior factors appear the users charisticaries based on the environmental and personal factors, which consider as outcome expectations or users' characteristics. Based on the main problem and the social cognitive theory, this study was guided by the following four research questions.

Research Questions:

1. ***What are the key sentiment characteristics of tweets related to Covid-19 posted during the 60 days before and the 60 days after declaration of the Covid-19 pandemic?***
2. ***What are the key discussion topics emerged from tweets related to Covid-19 during the 60 days before and the 60 days after the declaration the Covid-19 pandemic?***
3. ***What are the relationships between sentiment characteristics and discussion topics in Covid-19 related tweets during the 60 days before and the 60 days after the declaration of the Covid-19 pandemic?***
4. ***What are the differences between Covid-19 tweets during the 60 days before and the 60 days after the declaration of the Covid-19 pandemic in terms of the sentiment characteristics and discussion topics?***

Answers to these questions could provide valuable information to healthcare providers, government authorities, and policy makers regarding how the people reacted to the pandemic which it can help them on managing the Covid-19 pandemic. The 60 days before the declaration of the pandemic was chosen because the Strategic and Technical Advisory Group on Infectious Hazards (STAG-IH) at WHO held its first meeting on the novel coronavirus outbreak on January 10th, 2020 (WHO, 2020b), after which many different decisions were made in different countries, such as stay-home-orders, closing of borders, and closing of non-essential businesses (Gostin & Wiley, 2020), which had major impact on people's daily lives. Using sentiment analysis and

topic modeling techniques to comparing users' Twitter behaviors in parallel time periods before and after the declaration could shed light on how the policy makers' decisions affected Twitter users as a cross section of the larger population.

Research Objective:

The primary aim of this research is to compare the characteristics of Twitter users as inferred from their tweets during the 60 days before and the 60 days after the declaration of the Covid-19 pandemic. The overarching goal of this study is to help health care consumers, health care providers, governments, and policy makers appropriately manage responses to the Covid-19 crisis. The Covid-19 information gathered from the Twitter may play a fundamental role in governments' policies and patients' health care outcomes. The information that people create, share, and consume in online health communities on Twitter provides an important opportunity for researchers to understand people's and caregivers' needs and help them manage them during a pandemic. The implications of this study are both theoretical and practical.

The theoretical implications are related to the application of social cognitive theory to a comparison of Twitter users' behaviors during two time periods of the Covid-19 pandemic. The methodologies of sentiment analysis and LDA techniques utilized to explore Twitter communications focusing on the Covid-19 pandemic extend the application of the theory. The practical implications of this study comparison of Twitter users' characteristics before and after the declaration of the pandemic is that it sheds light on differences between users' concerns and behaviors before and after a full-scale pandemic has been declared and how they generate and share information and concerns on Twitter. It is hoped that the findings and the provision of a model for using social media to monitor relevant health outbreak will help policy makers deal with health crises now and in the future.

Chapter 2. Literature Review

In this chapter, the literature relevant to the topic of this study—the comparing users' characteristics on Twitter during early stage of Covid-19 pandemic—is reviewed. To our knowledge, limited study was conducted to compare sentiment characteristics and discussion topics before and after the declaration of pandemic by WHO, this review is focused on a broad description of health information on social media in related areas, such as health information, social media, users' behavior, data mining, previous health crisis, and machine learning. Accordingly, this chapter is divided into sections and subsections, the main ones of this which are as follows: health information on social media, data mining, and machine learning: sentiment analysis and topic modeling. These sections provide a research context within which to investigate the extent to which users' characteristics in Twitter effect before and after declaration of Covid-19 pandemic.

Health information on social media:

Overview:

In what is often known as the information age, there are various categories of information, such as health information, which is of broader interest than most other fields as health is important to everyone. The American Health Information Management Association defines personal health information as "data related to a person's medical history, including symptoms, diagnoses, procedures, and outcomes (*HIM Careers - Health Information 101*, 2019). According to the U.S. National Library of Medicine website, the broad category of health information has several sub-categories including general health, drugs and supplements, health of specific populations, genetics, environmental health and toxicology, clinical trials, and biomedical literature (Medicine). To ensure its quality, information disseminated by the

healthcare industry needs to have certain characteristics, including accuracy, precision, reliability, consistency, timeliness, relevance, completeness, comprehensiveness, availability, and accessibility. Health information that meets these criteria should be distributed consistently so that it may reflect positively on the healthcare industry.

Types of health information:

There are many different sources of health information, such as books, journals, government agencies, health organizations, the Internet, and social media networks. Social media have become more popular than others because of the nature of the information there. As Twitter is ranked as the third most popular social media network (Jain, 2012), it often plays an essential role in dissemination of health information on Twitter. Users and health organizations can publish what they want with few restrictions due to the democratic principle of free speech. However, Twitter's policy prohibits tweets with content that is considered harmful, such as violence, terrorism, child sexual exploitation, abuse, incitement to hateful conduct, self-harm, information that violates others' privacy, copyrighted or trademarked content, and anything that negatively affects election integrity (*The Twitter Rules*, 2019).

Main topics of online health information have been found to include specific diseases, symptoms, surgical procedures, daily health-related activities, medications, alternative therapies, and health systems (Underhill & McKeown, 2008). People often discuss personal health information on many different platforms, including Twitter, investigated in this research. According to Sokolova et al. (2012), the large number of people sharing their personal health information (PHI) online has created a need for the development of tools that can gather and analyze such information. There are two semantic-based methods for mining PHI, WordNet, which is used as a source of health-related knowledge, and search terms for personal relations.

The exchange of personal health information varies and is often shared from a healthcare perspective. According to Zhang et al. (2014), information exchange behavior in social media is different from general health information disseminated through non-social media because it is connected specifically to the physical and mental health of users, whose primary motivation is to gain knowledge and social value by being part of healthy communities. (Cain et al., 2000) identified three primary groups using the internet to obtain health information: those seeking general health-related information, the chronically ill, and the newly diagnosed seeking information about their condition. Health information is important to every person around the world, and Twitter provides the vehicle for exchange of health information.

Social media:

History:

Web 2.0 is an interactive version of the World Wide Web "WWW" and has a multitude of applications, such as email and social media. In 2004 Dale Dougherty, a vice-president of O'Reilly Media Inc, officially coined the term “Web 2.0’ (Andersen, 2007), and in (2007), Tim O'Reilly, the founder of the company, furthered this discussion with a famous paper entitled, “What is Web 2.0?” Several Web-based services and applications rely on Web 2.0, such as blogs, Wikipedia, Facebook, Twitter, Instagram, Snapchat, and LinkedIn. The Web 2.0 revolution has been regarded as opening a new path for democracy, allowing people to share their opinions or experiences in a public space. Following this revolution, social media now comprise top-rated applications used around the world that support users’ needs, such as communication in different languages.

Merriam Webster (2019) defines social media as, "forms of electronic communication (such as websites for social networking and microblogging) through which users create online

communities to share information, ideas, personal messages, and other content (such as videos). Social media networking has become one of the most efficient and cost-effective avenues for reaching a large, targeted audiences. It provides a powerful method for attracting followers and supporting interactions essential in various aspects of life. Social media networks enable users to create, modify, discuss, and share information in a network environment (Lober & Flowers, 2011) and have become the most popular platforms for people to communicate with each other. Grajales et al. (2014) describe the major types of social media services:

1. Blogs: They allow new users in the lay public to create a communal website, such as Wix, where opinions about any topic could be voiced, which creates communal, collaborative dialogues. This is the first (Web 2.0) social media type, created in the late 1990s.
2. Microblogs: They allow users to view a large number of updates with brief content over a short period of time. A microblog, such as Twitter, is considered the most dynamic and concise form of information exchange. Moreover, Twitter has remained the most prominent service on the market for these kinds of updates.
3. Wikis: They allow users to quickly and easily publish and edit articles.
4. Mashups: These are combinations of two or more Web services that use application programming interfaces (APIs) to create a new service or functionality, such as Google Earth.
5. Collaborative Filtering: These websites allow multiple users to tag or classify and crowdsource information to create a user-based, bottom-up folksonomy, which can be found in most blogs (e.g, to classify blog posts into one or more subject or theme), microblogs (e.g, through the use of hashtags), and wikis (e.g, to find related articles).

6. Media Sharing Sites: These are optimized for viewing, sharing, and embedding digital media on other Web services, such as YouTube.
7. Multi-User Virtual Environments: These allow users to interact with each other through a virtual representation of themselves known as an avatar, such as those seen in most video games.

There is a connection between social media life and real life in how people interact with information. According to Dietz-Uhler & Bishop-Clark (2001), most people who efficiently manage their online relationships also have good connections in their real-world lives. Soares (2012), in a study of the role of social media in supporting social relationships, found they play an essential part by increasing the speed and convenience of engagement among users, and that . people trust the health information given on social media differently from other information.

There are variations among social media platforms. According to Chu (2011), based on their different features, social media platforms can be classified into many different types depending on their applications and technologies, such as social networking sites including Facebook and microblogs, Twitter and content communities, and YouTube. Lin and Chang (2018) investigated these features in data collected from Facebook users; however, more research on other social media types to obtain a broad view of the social media landscape. Different social media platforms provide an opportunity for users to choose based on their different purposes and needs.

Health information on Twitter:

While some researchers have regarded Twitter as just a general communication platform, others have created standard metrics for measuring users' Twitter behaviors including, but not limited to, the number of tweets, retweets, and/or followers (Vega et al., 2010). In 2019, there

there were more than 326 million monthly active users on Twitter (@JacquelineZote, 2019), who were tweeting more than 590 million times per day, sending an average of 6,000 tweets every second (*Twitter Usage Statistics - Internet Live Stats*, 2019). Consequently, Twitter is considered the largest microblogging service. Using Twitter to increase healthcare awareness can have a significant impact on people's lives. According to Castillo et al. (2011), health organizations can utilize Twitter to promote health literacy, increase their name recognition, and manage their reputations. The researchers also discovered that about a third of health tweets were re-tweeted.

Exchanging health information on social media networks can change people's lives. (Lin & Chang, 2018) assert that studying the predictors of health information exchange can enable researchers to gain insight into how to utilize health information. Additionally, they found that understanding interaction (both person-to-person interaction and human-to-information interaction) and results for expectations of health self-management, efficacy, and social relationships can positively influence health information exchange.

Health information on Twitter can be spread quickly and widely, as users pay attention to health information by tweeting and interacting with each other. A hashtag, which can be used to point to tweets that are relevant to particular topics, is one method that facilitates discussion and increases the number of readers or followers. Hashtags enable Twitter users to discuss topics of common interest, including healthcare, and health-related hashtags constitute one of the characteristics of Twitter that serves as a quality source of health information retrieval.

Characteristics of health content on social media:

Availability and accessibility:

Social media such as Facebook, Twitter, YouTube, and blogs have a unique characteristic as a health information source in their capacity to provide accessibility to any users at any time

and location while other sources may not be as easily accessible (Kim et al., 2014). For instance, libraries have certain operating hours, which makes their information sometimes inaccessible. Thus, social media enable users to communicate and collaborate creating massive information spaces without location and time constraints, which differentiate them from traditional sources (Lai & Turban, 2008). Because availability and accessibility enable users to continuously post on social media, they rapidly proliferate health information.

Content Creation and Rapid Growth:

From the early 2000s, as more and more users participate in content creation rather than only consumption, user-generated content on the Internet has become increasingly popular (Agichtein et al., 2008). In Web 2.0, user-generated content continues to contribute to social media domains, including blogs, video sharing communities, and social networking platforms, such as Facebook and Twitter, which offer a combination of all of these affordances with an emphasis on the relationships among the participants in a community (Agichtein et al., 2008). Accordingly the growing use of social media facilitates peer-support, information creation, and sharing among people managing long-term health conditions (Hunt et al., 2015). Rapid growth is one of fundamental characteristics of social media as sources of health information retrieval. For instance, communications regarding chronic health issues have sped up the pace of increasingly high levels of user-generated content and formation of vibrant online communities (E Hilliard et al., 2015; Yonker et al., 2015). Consequently, content creation rapidly increased the generation of data to constitute what is referred to as Big Data.

Detecting Trends:

Based on its monitoring of global media sources and news websites, The Global Public Health Intelligence Network (GPHIN) has stated that epidemic outbreaks account for

approximately 40% of the World Health Organization's (WHO) early warnings (Mykhalovskiy & Weir, 2006). Social media could provide alternative low-cost public health indicators that serve as the basis for public health surveillance systems. In some circumstances, professional health institutions manage the public health surveillance systems that require regular clinical reports and considerable effort by health professionals to analyze data (Mykhalovskiy & Weir, 2006). For instance, the Center for Disease Control and Prevention (CDC) gives early alerts of health crises that typically incur a one to two week reporting delay (Ginsberg et al., 2009). Certain events, such as breaking news, health crises, and innovations attract the attention of social media users, whose volume of

responses can create trends based on health crises or other issues that enable researchers to understand the current situation globally and expand characteristics of social media as a source of health information retrieval. Moreover, health information seeking behavior on social media has unique characteristics that can provide insight into the characteristics of social media and users' behaviors.

Characteristics of Users' behavior:

Health information seeking behavior:

Olenja (2003) defined health information-seeking behavior (HSIB) as "any action undertaken by individuals who perceive themselves to have a health problem or to be ill for the purpose of finding an appropriate remedy" (2003). Among various methods for seeking health information, in this study the focus was on online HSIB, such as the Internet and social media. Lambert and Loiselle (2007) analyzed the concept of health information-seeking behavior from Wilson's (1963) conceptualization to the time of their study and found that "explicit definitions

of HISB are difficult to locate, and there is no apparent dominant definition" (p. 1008). Different definitions of health information-seeking behavior are shown in Table 1.

Table 1. *Definitions of Health Information-Seeking Behavior*

Authors	Definitions
(Lenz, 1984)	"Series of interrelated behaviors that can vary along two main dimensions: (a) extent and (b) method" (p. 63)
(Barsevick & Johnson, 1990)	"Actions used to obtain knowledge of a specific event or situation" (pp. 3-4)
(Corbo-Richert et al., 1993)	"Verbal or nonverbal behavior seeking to attain, clarify, or confirm information" (p. 30)
(Loiselle, 1996)	"A self-regulatory strategy that patients use to organize transactions between the self and health-related settings with the goal of balancing instrumental benefits and subjective costs stemming from informational outcome" (p. 9)
(Conley, 1998)	"Verbal or nonverbal behavior used to obtain, clarify, or confirm knowledge or information about a specific event or situation" (p. 132)
(Rees & Bath, 2000)	"Problem-focused coping strategy sometimes adopted by individuals as a response to a threatening situation" (p. 72)
(Rees & Bath, 2001)	Monitoring: "the urge to confront oneself with the threatening situation by means of seeking more information about it" (p. 900). Blunting: "tendency to distract from threat-relevant information" (p. 900).

Two fundamental dimensions of health information-seeking behavior are (a) the information dimension, comprising the characteristics of the information sought, particularly in terms of type and amount, and (b) the method dimension, which is related to the content and variety of the search (Lambert & Loiselle, 2007). The method dimension is also related to how much information about a given topic one seeks, indicating the depth of the search. Furthermore, the method dimension of HISB includes the optional actions individuals use to obtain health-related information and the sources from which information is sought, such as social media.

Among the factors that affect individuals' health information-seeking behaviors are socio-economic status, gender and its associated social status, age, availability of information on specific symptoms or diseases, accessibility of information, and perceived quality of the service (Tipping, 1995). In addition, MacKian classifies influences on health-seeking behaviors as geographical, social, economic, cultural, and organizational based on previously identified determinants in the relevant literature (2003). Thus, health information-seeking behavior depends on both internal and external factors related to individuals, such as education and family, some of which some can be managed. In an early study, (Gollop, 1997) reported that health information-seeking behavior related to age, education, literacy, and accessibility had a positive relationship to use of the library for seeking such information. In general, health information-seeking behavior is a critical step for individuals before going to providers.

The type of health information sought is closely tied to specific needs. According to the Health Information National Trends Survey (HINTS, 2019), 45% of information-seeking Americans have searched for cancer information, the most commonly sought topic across various sources. Several studies have been focused health information-seeking behavior on the Internet or social media (Basch et al., 2018; Kratzke et al., 2013; Park & Park, 2014; Rees &

Bath, 2001; Teufel-Shone et al., 2015). Social media are particular interest as reliable health information sources should be investigated because they can deliver information faster than other sources.

Users' behavior on social media:

Investigating users' information-seeking behavior should be the first step toward understanding how the users interact, search, and communicate through social media. User behavior have been investigated in the context of intensity of use, privacy concerns (Asiri et al., 2017; Zimmer & Hoffman, 2012), disclosure rates (El Ouirdi et al., 2015), personality traits, cultural norms, self-presentation, gender, age, and self-esteem (Waheed et al., 2017). Heinonen (2011)divided social media activities into three categories depending on the motivation for an activity: information processing, entertainment, and social connection. For information processing the user has to retrieve product information or content, collect factual information, share and access opinions, evaluate reviews and ratings, review images caught on surveillance cameras, surveillance of news , and application of prior knowledge (Heinonen, 2011).

(2017) point out that some factors can positively or negatively impact users' behaviors on social media, such as trust, regard for privacy, age, culture, gender, distance from source, and information sharing. They also identify important characteristics of social media users' behaviors such as frequency of use, information control, self-disclosure, self-censorship/self-awareness, self-efficacy, ease of use, self-presentation, social affiliation, social connection, social capital, personality traits, attitude, gratifications, relaxation, self-esteem, social presence, social influence, social norms, surveillance, regrets, emotions, boredom, reciprocity, shyness, and self-control. They also classified characteristics that directly affect social media users' behaviors into social investigation, social affiliation, frequency of use, information control, self-orientation,

reciprocity, and social boldness. As social media platforms, including Twitter, have different characteristics from each other, users' behaviors are highly changeable.

With regard to why people use Twitter, what users' intentions are, and what people tweet, Java et al. (2007) described a user active if he/she has posted at least one post during a week, and retained if he or she has reposted at least once in the following week (2007). They described the following four primary types of postings on Twitter (2007):

1. **Daily Chatter:** Most posts on Twitter concern the users' daily routines or what they are currently doing.
2. **Conversations:** To simulate the exchange of comments in a conversation, early adopters used the @ symbol followed by username to reply to others' posts.
3. **Sharing Information:** About 13% of all the posts in the research contained a URL to link to another site. Due to the small character limit, a URL-shortening service is frequently used to make this feature feasible.
4. **Reporting News:** Many users reported the latest news or commented about current events on Twitter. Some automated agents post updates such as weather reports and new stories from RSS feeds. Java et al. (2007) also posit the following main categories of users on Twitter based on the link structure of their postings:

1. **Information Source:** An information source is a hub or user who has a large number of followers. This user may post updates regularly or infrequently.
2. **Friends:** Most online relationships fall into this broad category. There are many sub-categories of friendships on Twitter.
3. **Information Seeker:** An information seeker is a person who might rarely post but follows other users regularly.

Data Mining on Twitter:

Overview:

Hand et al. (2001) defined data mining as "the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" (p. 6). Earlier, Cabena et al. (1998) posited that

> "Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases" (1998). Most recently, Witten et al. (2016) defined data mining as, "the process of discovering patterns, automatically or semiautomatically, in large quantities of data— and the patterns must be useful" (2016, p. 8)

The massive growth of social media usage has created a rapidly increasing accumulation of data, referred to as "Big Data." Compared to other data sources, social media have the unique feature an essential source of "Big Data." The nature of the data is not entirely new to the social sciences, but social media data inherently creates particular challenges, which Williams et al. (2017) referred to as the "6 Vs, and presented them below:

1. Volume: The massive amount of data being generated on social media platforms.
2. Variety: The different types of data including text, images, videos, geospatial check-ins, and audio.
3. Velocity: How the speed at which social media data are generated and quickly users respond to real-world incidents.
4. Veracity: The accuracy, reliability and overall quality of the data.
5. Virtue: Ethical treatment of data and the individuals who generate them.

6. Value: Refers to the evaluation of how social media data growing our understanding of the social world.

The rapid growth of social media data provides new opportunities for analyzing many aspects of and patterns in communication in different fields. Many previous studies are case studies (Achrekar et al., 2011; Qasem et al., 2015; Valli et al., 2017; Wang et al., 2018) that provide in-depth data on a specific subject for a specified amount for quantitative analysis. Social media analytics combine, extend, and adapt methods for social media data analysis (Stieglitz et al., 2014). Different types of data may be collected from social media, including textual and multimedia content, user profile pages, network data, and tracked activities such as likes (Sloan & Quan-Haase, 2017).

According to Fan and Gordon (2014), social media analytics have three primary steps: "capture," "understand," and "present." "Capture" is the act of obtaining relevant data by attending to different social media platforms and storing the relevant data. (Fan & Gordon, 2014)"Understand" entails choosing data for modeling while removing noisy low-quality data and using different advanced data analytic techniques to develop a meaningful picture. "Present" involves understandably communicating the results of the first two steps. Data can be collected by an application programming interface (API) or through third party tools that offer ready access using the API (Sloan & Quan-Haase, 2017); however, some researchers prefer to collect data manually. Twitter generates rich data that can be utilized to understanding users' behaviors or content from different perspectives. Twitter's data, in particular, have special features that affect data collection and data analysis, which may have a positive or negative impact on these processes.

As a single tweet has many attributes, there are various ways to gather data from Twitter. Three important points that should be considered at the outset of data collection are accuracy, reliability, and privacy (Alfantoukh & Durresi, 2014)., The health information data collected from Twitter could help predict the health issues that might accompany the spread of Covid-19 throughout the world and facilitate identifying the symptoms, which might inform the strategies the improvement of health organizations in order to reduce the spread of a disease (Valli et al., 2017). For example, Byrd et al. (2016) reported that the spread of influenza was traced by collecting the information from the tweets. Health information-seeking on Twitter can provide a useful tool for health organizations to build solid strategies.

Many researchers have based their investigations on Twitter communications, signifying the power of Twitter's data. Twitter provides real-time notification of unfolding events, such as the spread of diseases, the occurrence of natural disasters, political issues, live traffic updates, and business updates (Valli et al., 2017). By collecting tweets containing the keyword "flu," Achrekar et al. (2011) predicted the emergence and spread of influenza among the target population. Lampos and Cristianini (2010) also used Twitter data to fund correlations between the Twitter and HPA flu rates to calculate a flu score. These studies demonstrated the usefulness of Twitter data for gaining insight into the incidence and spread of common diseases and suggests the value of extending this approach into different areas.

Twitter is a very practical tool for mining large amounts of data for analysis and interpretation. The results can be used to in various ways, such as raising health consciousness and creating a long-term disease detector. An Application Programming Interface (API) can help the researcher to accumulate and analyze Twitter data to build an accurate model of the spread of a disease and its effects on a society (Hirose & Wang, 2012). According to Smith (2017), 95% of

top-ranked hospitals use social media for engaging with patients and providing accurate and credible information, which increases the user's trust and leads to his/her returning to the platform for social support, which increases the validity of information the platform accumulates (Hajli, Sims, Featherman, & Love, 2015).

According to Al-Taee et al, (2012), social media platforms have been used to create awareness of different diseases. Patients seek health information and find emotional support on social media, which facilitate connects among patients suffering from similar diseases. While many innovative technological solutions support patients during treatment, social media play a unique role by fostering interactions among people coping with the same health issues. Technologies can have a significant impact on provision of healthcare services around the world and has the potential of influencing social, cultural, and economic contexts (O'Connor et al., 2016). Developing technologies can enhance the health information shared on social media and address the queries of different users around the world.

CRISP-DM:

CRISP-DM stands for Cross-Industry Standard Process for Data Mining, which is used as a methodology and process model (Shearer, 2000) and is the most widely used analytics model. CRIS-DM facilitates control of a data analytics project and maintenance of the quality of analysis. Figure 2. illustrates the process.

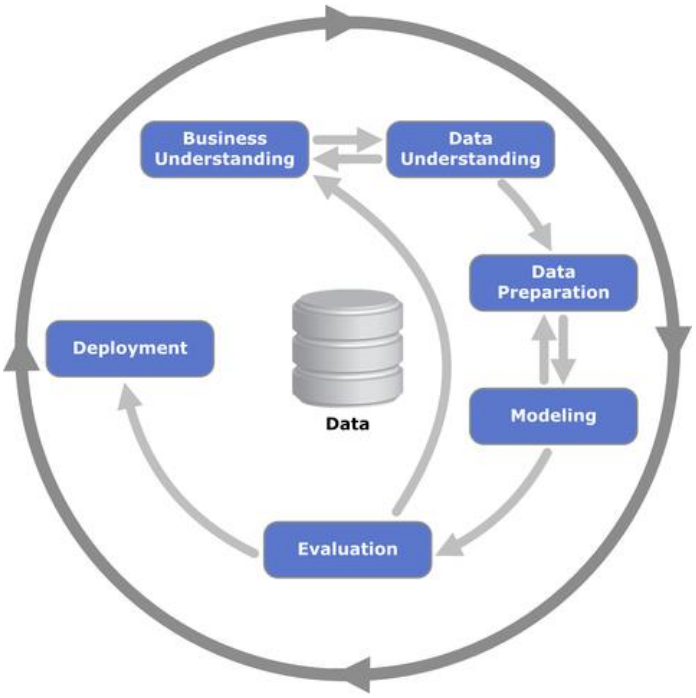

Figure 2. CRISP-DM

Health Organizations:

(Neiger et al., 2013) used Twitter to investigate sharing information, engaging with followers, and promoting action in local health departments (LHDs)as well as to discover the variations in Twitter use among them according to the size of population served. They collected and analyzed 3,000 tweets from LHSs and found that 56.1% were related to personal health concerns and 39.5% to the organization. They also discovered that small LHDs were less likely to tweet about personal health recommendations for individuals to take action to change their lifestyles, while people in large LHDs were more likely to tweet about engagements with their organizations and less likely to suggest that users take action that would benefit the organization. The researchers concluded that as Twitter was being adopted by LHDs to interact with their audiences in conversations, its main use was for one-way communication on personal-health topics with individuals as well as organization-related information.

Han et al. (2019) used a synthesized analytical approach to textual analysis to examine how key organizations and individuals used Twitter strategically to spread health messages and functions. They found that key health providers used Twitter for spreading health information, building relationships, or encouraging people to take health-related actions, Healthcare organizations associate with other businesses in order to connect with users through Facebook, Twitter, and other social media platforms thus empowering consumers to manage health crisis (Lober & Flowers, 2011). For instance, Twitter was used as a real-time method for managing the surveillance of the Ebola outbreak in order to monitor the spread of information and achieve early epidemic detection (Odlum & Yoon, 2015).

George et al. (2018) examined 812 posts, 507 (62.4%) of which were related to the study, to evaluate the characteristics of social media hashtag campaigns relevant to health and compared three different campaigns on Twitter, Instagram, and Facebook, #let's talk (World Health day 2017), Hands up #HIV prevention (World AIDS day 2016), and #No tobacco (World No Tobacco Day 2017). They also compared the sources, publicity efforts, credibility, reach-outs, and other characteristics of the three social-media campaigns. They concluded that posts on social media related to hashtag campaigns are highly credible and relatable, but not highly popular although users can use hashtags to seek health advice. Nevertheless, health credibility is a serious issue in online health communities (Briones et al., 2012). Thus, healthcare providers and organizations may be trustworthy sources of advice on healthy lifestyles and medical tips.

In exploring the characteristics of social media as a source of health information. retrieval, health information seeking behavior (HISB) and users' behaviors on social media are crucial elements to understand. Seeking behaviors can differ based on the type of information sought. For instance, business information is often compressed into figures while health

information usually has more details. Recently, healthcare providers have paid close attention to social media as its consumers consider them a reliable source. Healthcare organizations target improving consumers' health by using social media platforms to publish health information (Koh & Tan, 2011; Prybutok et al., 2014).

Previous Health Crisis:

Kostkova et al. (2010) presented their analysis of over one million tweets discussing the H1N1 (swine flu) pandemic from May to August 2009 and discovered many notations of specific words and phrases related to the H1N1 outbreak, including 2,888 instances of "I have swine flu" and 1,530 occurrences of "I have flu." Ritterman et al. (2009) collected 48 million tweets during the three months to predict locations of H1N1 flu outbreaks. Both studies suggest that further investigation is needed into the possibility that microblogging can be leveraged to better understand public knowledge of infectious outbreaks and to create early warning systems.

Michelle and Sunmoo (2018) evaluated Ebola-related health information needs during the peak of the pandemic via longitudinal tracking by collecting and analyzing 155,647 tweets posted from July 2014 to March 2015 by applying NLP and content analysis techniques. Their findings show that a shortage of health information caused fear and frustration, and social media did not serve the health information needs of individuals. In addition, Liang et al. (2019) examined the following and retweeting patterns Ebola information on Twitter by collecting all Ebola-related tweets posted around the world from March 23, 2014, to May 31, 2015, and classified users into four categories: influential users, hidden influential users, disseminators, and common users. They found 91% of the retweets were directly from the initial tweet. They believe that public health communicators can cooperate beneficially with influential and hidden

influential users to disseminate information, because these users can spread tweets to reach more people who are not following the public health Twitter accounts.

Edin et al. (2016), in a systematic review of social media effects on patients and on their relationships with healthcare professionals, identified six categories of patients' uses of social media: emotional, informational, esteem, network support, social comparison, and emotional expression. These studies suggest that patients' use of social media for acquiring health information should be investigated further. The content of health information conveyed by social media has unique characteristics compared to traditional sources.

Machine Learning: Sentiment analysis and Topic Modeling:

Overview:

Machine learning is based on the assumption that programs should have the ability to learn from training data. It is a technique of data analysis involving automated analytical model building. Machine learning has developed in response to two interconnected questions: How can one construct computer systems that automatically improve through experience? and What are the fundamental statistical computational-information-theoretic laws that govern all learning systems, including computers, humans, and organizations? (Jordan & Mitchell, 2015). Machine learning is an Artificial Intelligence AI) techniques that is widely used in many fields. Machine learning is essential both for pursuing fundamental scientific and engineering questions and for generating highly practical computer software fielded across many applications. There are many different types of machine learning, such as image recognition, speech recognition, medical diagnosis, predictive analytics, and text analytics. Sentiment analysis and topic modeling are among the most popular machine learning techniques and are widely use in text analytics.

The aim of sentiment analysis is to detect meanings behind words by analyzing people's opinions, attitudes, reactions, and/or emotions as expressed through texts, such as consumer reviews, survey responses, and social media (Liu, 2012). The democracy of social media platforms encourages people to use them to express their opinions. Many websites such as Amazon.com welcome people's written opinions about products. On Twitter, users can write their opinions on various topics. Sentiment analysis is employed in many areas, such as marketing, product reviews, emotion detections, etc. Various techniques are used, such as natural language processing, computational linguistics, and other machine learning algorithms.

Over the past 20 years, many researchers have conducted sentiment analyses using different techniques of sentiment classification. Webster and Kit (1992) explain that the main step of sentiment classification is tokenizing the target text into elements such as word-entities, character-entities, etc. Different machine learning methods use supervised or unsupervised algorithms or both along with different features. Supervised learning techniques are used with labelled classes of a dataset, and it is recommended that the text be cleaned of punctuation, numbers, etc. to avoid noise in the dataset and improve the classification (Haddi et al., 2013). This pre-processing of the text can improve the result of the sentiment analysis by removing stop words, removing numbers, stemming, and lemmatizing the text, etc. After cleaning the text, researchers can select or extract features from the tokens, using n-grams, part-of-speech tagging, and bag-of-words models, which will be converted to numeric values using different methods, such as Term Frequency (TF), Term Frequency Inverse Document Frequency (TF-IDF), etc. Consequently, these numeric features become inputs for employing sentiment analysis using supervised machine learning methods. On the other hand, unsupervised learning techniques, which are used with an unannotated dataset, can estimate sentiment scores using different

methods employing the format of a lexicon or a dictionary, in which the sentiment words or phrases have sentiment values. The lexicon mainly used for sentiment analysis is Affective Norms for English Words (ANEW) (Bradley & Lang, 1999). Linguistic Inquiry and Word Count (LIWC) is an alternative method that was started in 2001 and further developed in 2007 and 2015 (Pennebaker et al., 2015). Both will be explained more in the methodology chapter.

Previous Sentiment Analysis and Topic Modeling Studies:

Roberts et al. (2012) discovered seven types of emotions on Twitter, which are anger, disgust, fear, joy, love, sadness and surprise. The data were collected from a variety of topics on Twitter, including Valentine's Day, Lindsay Lohan, September 11th, the 2012 U.S. Election, Palestinian statehood, the Egyptian riots, Super Bowl XLV, the 2010 World Cup, Christmas, the DC/NY earthquake, the Emmys, Eminem, the stock market, and the Greek bailout. They classified the texts using the Support Vector Machine (SVM) classifier and different classification features including unigrams, bigrams, trigrams, WordNet synsets, WordNet hypernyms, topic scores and significant words. They then used a 10-fold cross method to validate the classifier model. They used various methods to get a highly credible result, but the diversity of topics led to a variety of non-sentiment words, which made noise in the features and the dataset.

Go et al. (2009) applied three machine learning algorithms to the emotional datasets that were collected from Twitter using extractions of word features such as unigrams, bigrams, combinations of unigrams and bigrams and combinations of unigrams and parts-of-speech tags. Cleaning the texts of usernames, links, and repeated letters in tweets deflated the original texts to 45.85%. They also used a technique to map the emotions, such that :), :-), :), :D, and =) were all mapped to :). They achieved the best classifier's accuracy, which was 83.0%. They deleted all

tweets that had both positive and negative emotions, which could create a third class of mixed opinion.

Kumar and Sebastian (2012) demonstrated a new technique of sentiment analysis for Twitter, which involved extracting opinion words in the form of adjectives, verbs and adverbs, using corpus-based and the dictionary-based methods to find the semantic orientations of adjectives, verbs and adverbs. After finding the opinion words, they implemented scoring modules for the adjective group and the verb group. Lastly, they calculated the tweets' sentiment score using a linear equation. The limitation of this approach is that only adjectives, verbs, and adverbs are considered. Moreover, most Twitter users use irregular spellings, abbreviations, and slang, such as "besty," which means "best," which makes recognizing and classifying words for sentiment analysis purpose difficult.

Agarwal et al. (2015) applied five methods (baseline, domain-specific ontology, the importance of the feature, contextual information, and a combination of context information and the importance of the feature on three different datasets: restaurant reviews, software reviews and movie reviews. They divided the datasets into 90% for training and 10% for testing and found that the combined method, with accuracies of 80.1% for the software dataset, 78.9% for the movie dataset and 79.4% for the restaurant dataset, was the best. The advantage of this study was the use of ontology to reveal the entities and aspects in the tweets. However, the study relied only on the values of the words given in the lexicons and the ontologies' relations among these words. Consequently, the values of the words in the lexicons would affect the sentiment scores. Additionally, Mamidi at el. (2019) analyzed public sentiment concerning the Zika virus using tweets and determined sentiment characteristics using machine learning techniques and algorithms. They demonstrated how sentiment expressed within discussions of epidemics on

Twitter, which were mostly negative, could be discovered and enable public health officials to understand public sentiment regarding an epidemic and address specific elements of negative sentiment in real time.

Zhang et al. (2011) applied sentiment analysis a Support Vector Machine (SVM) algorithm to classify five different datasets: Obama, Harry Potter, Tangled, iPad, and the Packers. They found the average accuracy of the method was 85.4% and the average F-measure = 74.9%, precision = 68.7% and recall = 82.7%. They applied only unigram binary feature values with attention to negations, but other feature extraction methods should be used to reveal the strength of their model.

Jianqiang and Xiaolin (2017) investigated the impact of applying six methods of pre-processing the text on the performance of sentiment classification, which included removing URLs, removing numbers, removing stop words, normalizing repeated letters, normalizing acronyms to their originals, and normalizing negative mentions. They employed five datasets to evaluate the use of four classifiers. They indicated that removing numbers, stop words, and URLs reduced the noise in the datasets, while normalizing negative words and acronyms improved the performance of classification. The researchers employed the sentiment analysis using only three levels, "positive", "neutral", and "negative." In the normalization stage, they did not normalize the repeated letters to the original forms, e.g., normalize the word "goooooood" to "good." This method effectively discriminated between the classes "positive" and "very positive," so the word "good" has a certain meaning and it can be classified as "very positive" whereas, the word "good" can be classified as "positive." However, the main problem with this study is that there was no "very positive" class in their sentiment classifications.

Many recent research has used Twitter as the data source to carry out different techniques in a variety of forms of applications, such as predicting political preferences (Stier et al., 2018), driving business growth beyond marketing (Hanah, 2019), and monitoring the infectious disease and public health crisis (Liang et al., 2019). When people create, share, and consume information on Twitter, they create valuable opportunities for researchers to understand the needs of people and caregivers in online health communities and learn ways to help manage a pandemic.

Covid-19 Sentiment Analysis and Topic Modeling on Twitter:

Many scholars have studied various relevant aspects linked directly or indirectly to Covid-19. There are many areas, such as health, the economy, the environment, education, and politics, to which researchers can contribute. In this section, I discuss studies related the topic of the present research, which is coverage of the Covid-19 pandemic on Twitter.

Jimenez-Sotomayor et al. conducted a qualitative study of tweets connected to COVID-19 and older adults, using content obtained from 18,128 tweets. After they analyzed 352 tweets randomly selected as a sample, they classified the tweets into six groups: informative, personal accounts, personal opinions, advice seeking, jokes, and miscellaneous. They concluded that 25% of the analyzed tweets contained ageist or potentially offensive content relating to older adults, and most of the tweets comprised personal opinions, personal accounts, and jokes. Moreover, Raamkumar et al. (2020), in an examination of COVID-19-related outreach efforts of PHAs in Singapore, the United States, and England, and the corresponding public responses to these outreach efforts on Facebook, found that fear of COVID-19was the negative sentiment most frequently expressed.

Al-Rakhami and Al-Amri (2020) analyzed the credibility of information related to the COVID-19 pandemic posted on Twitter. They integrated six machine-learning algorithms with

ensemble learning in an ensemble-learning-based framework for verifying the credibility of tweets by analyzing a large dataset of tweets. They classified the data into credible and non-credible tweets and found that 70.22% of tweets were non-credible, and 29.78% were credible. They conclude that their final model worked slightly better than the original model, and the size can include fewer features.

Gupta et al. investigated Twitter users' awareness of the weather's impact on the spread of COVID-19 by performing natural language processing and machine learning methods (2021). They collected over 166,000 tweets between January 23 and June 22, 2020, and applied ML/Natural Language Processing (NLP) methods to screen for related tweets, classify them by the type of effect they claimed, and identify topics of discussion. Their results showed 28,555 related tweets, of which approximately 40.4 % indicated uncertainty about the weather's effect, 33.5 % indicated no impact, and 26.1 % indicate some impact. The researchers concluded that social media platforms can be utilized as a mechanism to crowdsource topics for research in order to address public misconceptions.

Abd-Alrazaq et al. (2020) aimed to discover the main topics connected to COVID-19 posted by Twitter users in English from February 2, 2020, to March 15, 2020 by applying Latent Dirichlet Allocation (LDA) for topic modeling. As a result, they identified 12 topics, which were classified into four main themes: the origin of the virus; its sources; its impact on people, countries, and the economy; and ways of mitigating the risk of infection. This result of the mean sentiment yielded 10 positive topics and two negative topics, which were related to deaths caused by COVID-19 and to increased racism.

Cinelli et al. (2020) conducted a comparative study on five different social media platforms (Twitter, Instagram, YouTube, Reddit, and Gab) during the Covid-19 pandemic

employing Partitioning Around Medoids (PAM) and using the cosine distance matrix of words in their vector representations as a proximity metric. They collected data from January 27, 2020, to February 14, 2020, and identified a set of topics for each social media. On Twitter, they found 21 topics: suspended flights and repatriation, economic impact, protection advice, prayers, requests for God's blessing, death tolls, infection rates, biological warfare, communist regimes, Huoshenshan hospital, comparisons with other viruses, Chinese wet markets, virus spreading, disease descriptions and symptoms, racism, and other. After modeling the spread of information, they found that whether information was reliable or questionable, it had similar dissemination patterns. These social media platforms are not similar usage because Twitter usually uses with text and YouTube is for video. So, this can make unbalanced results.

Regarding sentiment analysis, Lwin et al. (2020) examined more than 20 million tweets posted from January 28 to April 9, 2020 employing a lexical approach and the CrystalFeel algorithm to analyze four emotions that emerged, fear, anger, sadness and joy, and the reasons associated with them. They also created word clouds to find topics potentially related to emotions. Based on the word clouds, they suggested that fears were related to shortages of COVID-19 tests and medical supplies, anger shifted from xenophobia to stay-at-home notices, sadness was related to topics of losing friends and family members, and joy included words of gratitude and good health. They used only four keywords for data collection, and they suggested that future studies should further investigate sentiments by examining specific countries and expanding the scope to include other media platforms. Additionally, Chakraborty et al. (2020) analyzed 226,668 tweets posted between December 2019 and May 2020 using sentiments analysis and found that the numbers of positive and neutral tweets were high and the retweeted tweets were the most negative. Both studies provide useful results related to sentiment analysis.

Gupta et al. (2020) collected Covid-19 tweets over a longer period, from January 28, 2020 to July 1st, 2020 and analyzed using the CrystalFeel algorithm. They annotated each tweet with 17 latent semantic attributes linked to the previous 10 topics identified by the LDA algorithm and seven attributes related to the sentiments retrieved. They found that anger was the dominant emotion in the tweets. This study has one limitation related to data collection which they used only five keywords.

Garcia and Berton (2021) conducted a comparative study on Covid-19 between Brazil and the U.S using topic identification and sentiment analysis. They collected 3,332,565 tweets in English and 3,155,277 tweets in Portuguese between April and August 2020 to explore and compare strategies in both countries. They identified 10 primary topics related to COVID-19, of seven were equivalent in both languages, and a negative result of sentiment analysis in both languages related to proliferation of care, case reports, and statistics. They suggested that analyzing social network contents can give us a perception of society and the world. This study has one limitation that related to the keywords employed to retrieve content related to COVID-19. They suggested conducting more comparative studies related to Covid-19.

Latent Dirichlet allocation (LDA), which is a technique for analyzing a large set of texts, was the most commonly utilized topic model. The premise of LDA is that documents are represented as a distribution in which each topic is characterized by a word distribution. Lyu and Luli (2021) pursued two primaries aims, which were discovering the topics and the CDC's overarching themes derived from public COVID-19-relevant discussions about the CDC on Twitter, and providing insight into the public's concerns, such as the focus of attention, perceptions of the CDC's current performance, and expectations from the CDC. The dataset comprised 290,764 unique tweets from 152,314 different users posted from 03/11/2020 to

08/14/2020. The researchers performed LDA analysis and found 16 COVID-19 topics that the public linked to the CDC, which they classified into four themes: knowing the virus and the situation, policy and government actions, response guidelines, and general opinions of credibility. Using the LDA technique helped the researchers to think deeply about the data.

Mackey et al. (2020) conducted a study on detecting COVID-19-related symptoms, experiences with accessing testing, and mentions of disease in recovery in over 4,400,000 tweets collected starting from 03/20/2020. They used an unsupervised machine learning approach that included keywords that could be related to COVID-19 symptoms. By using the Biterm Topic Model (BTM), they analyzed the data into groups of tweets containing the same word-related themes, and then they extracted the tweets in these clusters, manually annotated them for content analysis, and assessed their statistical and geographic characteristics. The researchers concluded that although the study had some limitations, the value was in its innovative method using data mining in combination with modeling to sift through a large volume of unstructured data to detect and characterize potentially underreported cases of COVID-19.

These studies show that Twitter is considered one of the biggest data resources. Collecting and analyzing these data can help researchers understand people's behaviors could support authorities' efforts to control the pandemic.

Summary:

In this chapter the term "health information" on social media in general and Twitter in particular was defined. Then, various techniques for collecting and analyzing social media data related to managing the health crisis and how health organizations were dealing with it were described. Also, data mining in general and in Twitter in particular was introduced and reviewed. Finally, research related to sentiment analysis and topic modeling using machine learning

techniques in general and to Covid-19 sentiment analysis and topic modeling on Twitter was reviewed. No previous study comparing users' behavior 60 days before the declaration of the Covid-19 pandemic and 60 days after was found. In the next chapter, the methodology of the present study is described.

Chapter 3. Methodology:

Overview of Research Design:

In this chapter the methodology used in conducting this study of Twitter users' characteristics before and after the declaration of the Covid-19 pandemic is described including the target population and sample, data collection and management procedures, data analysis, and assessment of the validity and reliability of the process. The relevance of the results for health consumers, healthcare providers, governments, and policy makers are discussed, focusing on effective response to and management of public health crises.

The research population comprised all tweets related to the Covid-19 pandemic posted in two strategically selected time periods, the 60 days before (January 11th, 2020, to March 10th, 2020) and the 60 days after (March 11th, 2020, to May 10th, 2020) the pandemic was officially declared. The data were obtained from a third-party provider, Georgia State University's Panacea Lab (Banda et al., 2020). Data cleansing, considered the most challenging stage in data collection, was conducted using Python and RStudio, which facilitate cleansing big data without omitting valuable data and provide free tools to analyze and illustrate data using various visualization techniques. RStudio's software is described on its website as "… an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management" (*RStudio*, 2019).

Sentiment analysis, emotional analysis, and LDA were employed to analyze the data. Figure 3 shows the primary stages of the research process, which included collecting the tweets from the Panacea Lab, filtering and pre-processing the dataset, and applying selection of different features and machine learning algorithms for sentiment analysis and LDA.

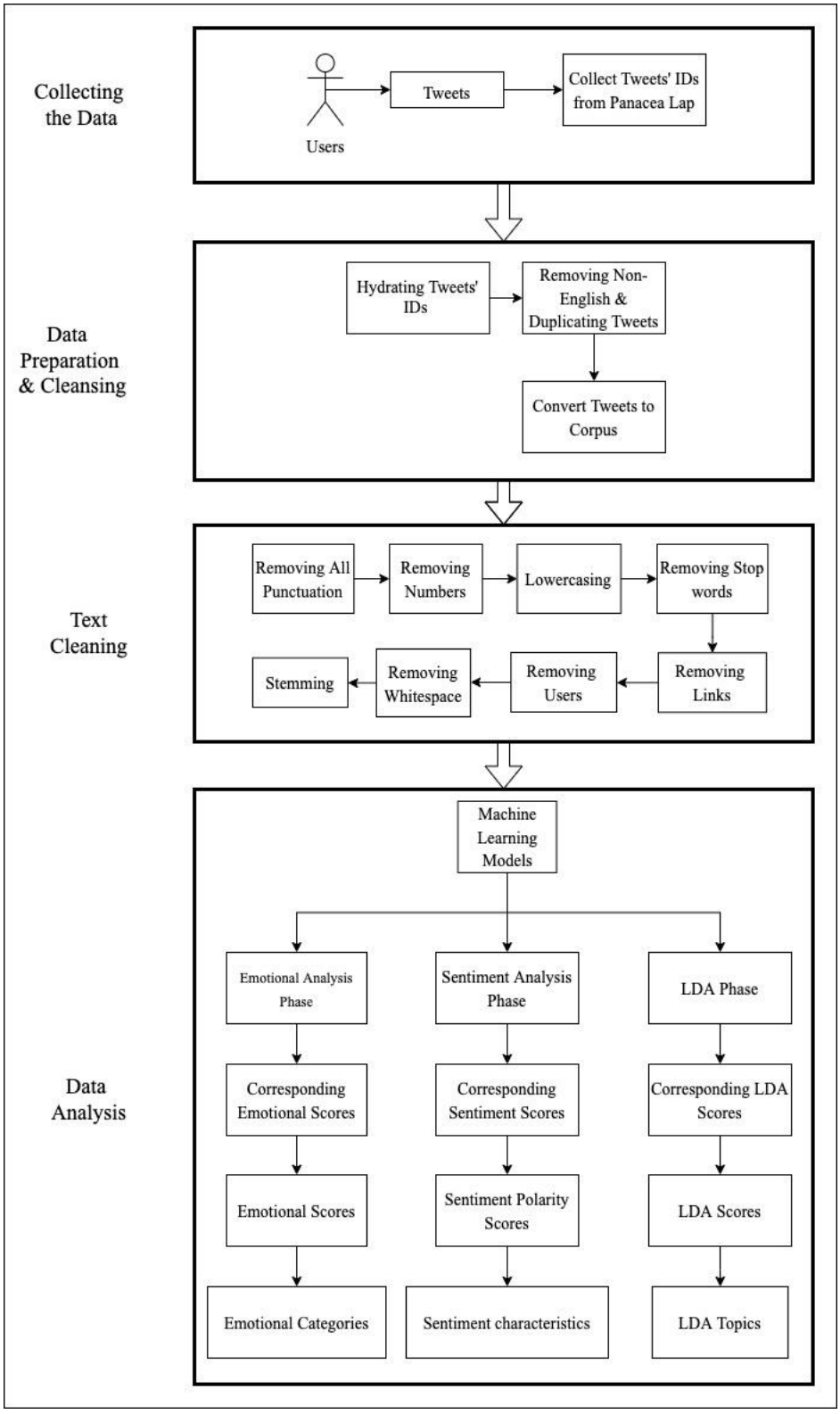

Figure 3. The Workflow of Data Analysis and The Main Steps

As noted, the focus of this study was on Twitter users' social media behaviors during the periods before and after the declaration of the Covid-19 pandemic. The objective of study was to compare the users' characteristics related to Covid-19 during these two periods to see how the declaration of the Covid-19 pandemic affect on users' characteristics. The Covid-19-related tweets included those posted by infected individuals, non-infected individuals, businesses, governments, researchers, health providers, etc.

The study population consisted of all tweets related to Covid-19, and the sample for this study included those posted during the two designated time periods and only English tweets. The total number of original tweets before excluding retweets and quoted tweets for this research, was 68,924,558. After retweet and quoted tweets were excluded, the total number was 21,655,284 tweets. These were categorized according to the research question, into tweets posted during the 60 days before the declaration of the Covid-19 pandemic (3,406,055) and tweets posted during the 60 days after the declaration (18,249,229). Due to the large volume of tweets, a data cleansing strategy was used to filter out garbage tweets. After the data were obtained from the third party provider, Georgia State University's Panacea Lab (Banda et al., 2020), each of the datasets was cleansed and prepared for analysis.

Data Collection Procedures:

To support researchers seeking insights into the implications of the pandemic since it started, Georgia State University's Panacea Lab has provided them with the opportunity to download Covid-19 related tweets (Banda et al., 2020). The Panacea Lab has systematically collected Covid-19 tweets since February 11th, 2020, first by using the keywords *coronavirus* and *2019nCoV* from January 1st, 2020 to March 11th, 2020. After that, they used a variety of keywords: *COVD19, CoronavirusPandemic, COVID-19, 2019nCoV,*

CoronaOutbreak,coronavirus , WuhanVirus, covid19, coronaviruspandemic, covid-19, 2019ncov, coronaoutbreak, and *wuhanvirus,* which produced around 4.4 million tweets every day. They provided only two attributes of each data item, the date and tweet ID.

Over 68 million tweets related to the beginning of the pandemic were posted between January 11th, 2020 and May 10th, 2020. Since the data will be collected via the third party, it needs more processing, such as hydrating and cleansing the data. To make the tweets readable, it was necessary to hydrate the tweet IDs by using hydrate application or Python to reveal other attributes, such as texts, number of likes, number of favorites, locations, languages etc. Analyzing such tweet content could reveal information about conversation topics, characteristics of users or organizations tweeting, and public opinions about specific topics.

Data Pre-processing:

Data pre-processing is normally a fundamental step before data analysis. There are critical steps involved in the data pre-processing phase, which are data modelling, data cleansing, and data transformation. These are explained below.

Data Cleansing:

After data have been collected and modeled, the next challenge is cleaning to prepare the data for analysis. Data cleansing is conducted to eliminate noise, inconsistencies, and errors that will be problematic later during analysis (Sloan & Quan-Haase, 2017) and can therefore improve the quality of the results. This step also involves correcting corrupt data, filtering irrelevant data out of the dataset, and streamlining the dataset. Because Twitter has a high volume of data, it may contain many irregularities. Sloan and Haase discuss four potential problems with data that may occur with social media data:

1. Missing data: The most common problems can occur at the example level, or because of a lack of (or overly permissive) validation within the social media application database schema. Researchers can avoid this problem by using a process called imputation, in which a variety of techniques are used to replace the missing data. This method uses machine learning to replace missing values and account for the imputation during analysis.
2. Data Entry Errors: These occur almost exclusively at the instance level because of individual user error, such as inconsistent spelling. Researchers can deal with this sort of error using a process called fuzzy string matching, which is the most common approach.
3. Duplicate Data: Duplication can appear in Twitter data, particularly data that have already undergone parsing or transformation. This problem can be dealt with after typos and spelling errors have been resolved, which often makes previously unrecognized duplicates visible within the dataset.
4. Inconsistent Units/Formats: This issue occurs when combining data from a variety of sources and can be dealt with based on the nature of data. In this study, however, Twitter was the only source of the data.

Data cleansing is an important step that can change a result in a positive or negative direction. Some software provides excellent tools for cleaning high volume data, such as RStudio, open source Apache Hadoop for processing, and text analysis with Python (Kim et al., 2013). However, these entail both monetary costs and learning efforts, which may be a deterrent for researchers conducting a one-time study.

Data Transformation:

After modelling and cleaning, data are ready for the final stage of data processing. In this step, data are transformed and consolidated into forms to which text mining and machine learning techniques may be applied. Sloan and Quan-Haase (2017) describe data transformation as "receiving input data and formatting it in a way that complies with a data model and can be imported into storage and analysis software" (Sloan & Quan-Haase, 2017, p. 138). This level of data processing consists of obtaining an input, extracting the relevant information by parsing the input, and outputting the information in a specified format. In this study data normalization processes were implemented to ensure that all the characteristics were shown at the same level, which is especially useful with the machine learning method and a standard range was used.

Tweets posts usually have different levels of language usage, often featuring colloquialisms, slang, abbreviations, spelling, and grammatical errors in tweets. In this study, traditional text processing techniques were applied, such as transfer of the text into corpus format; removal of stop-words (\the",\a", \an", \in", etc.), punctuation, the words "positive" and " negative" (because of positive and negative Covid-19 test results), unnecessary white spaces, and numbers using the "tm" library available in the RStudio programming language; word-stemming (e.g., converting "play," "playing," "played," and "plays" to the root word "play"); and making all characteristics lower case. The data were then ready for analysis.

Data Analysis:

Text measures on Twitter:

The four primary types of data analysis techniques are network measures, text measures, sentiment analysis, and predictive analytics, each with different branches. Text measures are the most common techniques that researchers use. Finding topics under text measures and sentiment

analysis are the two primary techniques related to text analysis. Finding topics is done to detect trends and peaks in discussion activity by identifying frequent words or words related to one subject area, such as “Football” and “Sports” (Kim et al., 2013; Kumar, 2013). Finding how users interact with particular subject matter on Twitter can provide a bigger picture as to the kinds of topics, such as a health crisis, with which they are concerned.

The Tweets preparation processes:

Two primary types of text data are the output of standard text mining systems and user-generated textual data. Standard text mining processes are often applied to formally prepared documents. User-generated textual data, on the other hand, are found on social media platforms, Internet webpages, etc., which it may not be controlled. Consequently, Twitter data need to be prepared prior to analysis. In order to prepare the raw text for in-depth natural language processing, such as topic model training, a series of steps must be systematically followed, which commonly consist of selection, cleansing and preprocessing of the text (Liddy, 2000).

In this study, text mining packages by Python and RStudio were used in the text preparation process. In Python, nltk and other essential libraries for data cleansing were used. In RStudio, “twitteR", "ROAuth", "tm", and "NLP" were used The tm package utilizes a corpus as its main structure to manage the text documents and provides easy access to preprocessing and manipulation mechanisms such as whitespace removal, stop word removal, stemming, and lowercasing (Feinerer et al., 2015). The first step was to convert the tweets in the dataset to a corpus. Then, the text was cleaned by removing all punctuation, numbers, stop words (e.g. “the”, “is”), and whitespace; lowercasing; and stemming. Finally, to ensure the accuracy of the subsequent analyses, the corpus was cleansed and manually checked by randomly searching the text.

Sentiment analysis on Twitter:

Definition of sentiment analysis:

Sentiment analysis or opinion mining is a process used to gage public opinion about particular phenomena such as services and products. It automatically assigns a positive, negative, or neutral "sentiment score" (Kim et al., 2013; Kumar, 2013) to a piece of text and thus can show how users express their feelings toward specific topics, such as negative feelings about the effects of Covid-19. Sentiment analysis can be utilized as a tool of for strategic decision-making likely to generate positive outcomes (Kang et al., 2018). For instance, a governmental authority may evaluate the emotions of the community concerning measures prescribed to control the spread of infection during the Covid-19 pandemic. If the community shows negative feelings about the restrictions imposed, the authority may decide to loosen them. With the rise of social media platforms on which many users share their opinions publicly, the use of sentiment analysis is growing substantially. .

Sentiment Analysis Processes:

In this study, the Valence Aware Dictionary and sentiment Reasoner (VADER),[1] an entirely free open-source tool in Python that uses machine learning techniques is specifically attuned to the sentiments expressed in social media, was used for sentiment analysis. VADER was developed by Hutto and Gilbert (2014), and its effectiveness has been assessed according to 11 typical state-of-the-practice benchmarks, including the Affective Norms for English Words (ANEW), Linguistic Inquiry and Word Count (LIWC), the General Inquirer, Senti WordNet, and machine learning-oriented techniques that rely on the Naive Bayes, Maximum Entropy, and Support Vector Machine (SVM) algorithms.

[1] https://pypi.org/project/vaderSentiment/

Emotional Analysis Processes:

RStudio, an open-source language and environment software, was used for statistical computing and generation of graphs (*RStudio,* 2019). RStudio provides many packages that allow users to manipulate data, one of which is the "Syuzhet"[2] package, used for the extraction of sentiment and sentiment-based plot arcs from text. Syuzhet means "device," referring to a narrative structure or fabula in which events follow chronological order (Jockers, 2017). The Syuzhet package contains four sentiment dictionaries and, Jockers (2020) explains, provides a robust, but computationally expensive, sentiment extraction tool developed by the NLP group at Stanford. The default "Syuzhet" lexicon was developed under Matthew Jockers' direction in the Nebraska Literary Lab. The other lexicons were developed as follows:

1. The "afinn" lexicon was developed by Finn Arup Nielsen as the AFINN WORD DATABASE.
2. The "bing" lexicon was developed by Minqing Hu and Bing Liu as the OPINION LEXICON.
3. The "nrc" lexicon was developed by Mohammad, Saif M. and Turney, Peter D. as the NRC EMOTION LEXICON (NRC) and is used to calculate the existence of eight different emotion and the corresponding valance within text file.

There are three features of the syuzhet package to be used which are:

1. The open-source nature of the syuzhet package can be used freely for research purposes.
2. It provides transparency and reproducibility.

[2] https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html

3. It will provide sentiment scores for ten sentiment categories, while other packages provide either a single sentiment score, simply two sentiment scores, or three sentiment scores one positive, one negative, and natural.

Scores were generated for each dataset according to the frequency with which certain words appeared in each of the following ten categories: Anger, Anticipation, Disgust, Fear, Joy, Negative, Positive, Sadness, Surprise, and Trust.

Latent Dirichlet Allocation (LDA)

Blei, Ng, and Jordan (2003) introduced Latent Dirichlet Allocation (LDA) as a generative probabilistic model of a collection writings called a corpus for topic discovery. They explained that the main purpose of LDA is to consider documents as random mixtures of latent topics in which each topic is represented by its distribution over words. Griffiths and Steyvers (2004) explained that LDA assumes a latent structure composed of a group of topics, and the texts that appear in a paper reflect the particular set of topics. Figure 4 shows the probabilistic graphical model of the LDA model. Rehurek and Sojka (2010) implemented LDA using Python with machine learning packages to apply the Gibbs Sampling inference method. In this study, LDA model in "sklearn"[3] package in Python was used to reveal the discussion topics. The algorithm uses stochastic optimization to maximize the variational objective function for the LDA topic model, which only looks at a subset of the total corpus of documents each iteration, and thereby is able to find a locally optimal setting of the variational posterior over the topics more quickly than a batch Variational Bayes algorithm could for large corpora.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

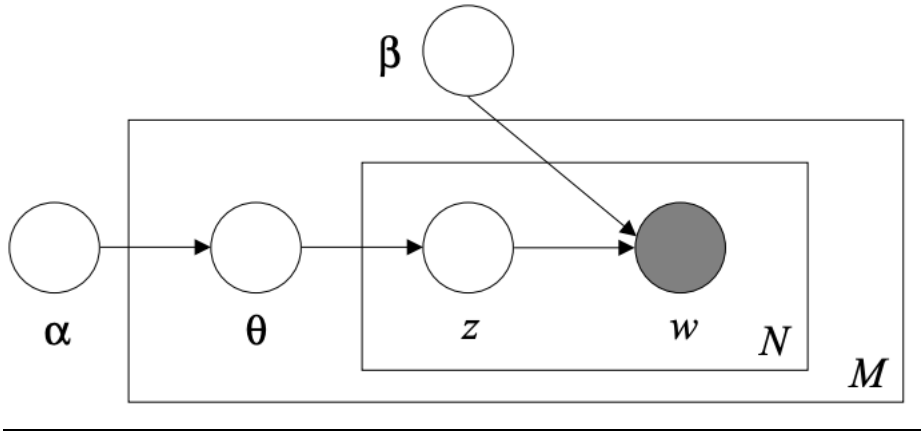

Figure 4. The Probabilistic Graphical Model of the LDA

Each node in Figure 4 is a random variable and is labeled based on its role in the generative process. The hidden nodes, which are the topic proportions, assignments, and topics, are unshaded and the observed nodes, which is the words of the documents, are shaded. The rectangles indicate “plate” notation, which is used to encode replication of variables. The 'N' plate indicates the collection of words within documents; and the 'M' plate indicates the collection of documents within the collection.

LDA Evaluation:

The performance of the topic model training is affected by the pre-specified numbers of topics. Cao et al. argued that LDA cannot produce correlations among the captured topics even though the topics discovered by LDA capture the correlations among words in documents. Fewer topics do not enable authors to be distinguished, while many topics may cause relationships to be weaker (Lu & Wolfram, 2012). Generally, topics identified in the documents are assumed to have differences from each other. Murdock and Allen (2015) introduced a way of evaluating the LDA model through interactive visualization, which supports rapid experimentation for interpreting hypotheses.

Inferential Analysis:

To test the differences between the two groups in sentiment characteristics and discussion topics, a series of inferential analyses was applied. Pallant (2013) suggests that parametric tests require the assumption that the shape of the population distribution is normal, while non-parametric tests require no assumptions about the underlying population distribution. To examine the normality of the datasets before proceeding with the parametric tests, a Kolmogorov-Smirnov, Mann-Whitney U, and Kruskal-Wallis H tests were conducted.

Validity and Reliability:

Validity:

Validity and reliability are the most notable criteria for evaluating the quality and appropriateness of the methodology of a study. Babbie (2016) defines validity as "the extent to which an empirical measure adequately reflects the real meaning of the concept under consideration" (2016, p. 534). To ensure the validity of this research, various data analysis techniques were applied to the data collected from Twitter to answer the research question. Twitter has unique features in comparison to other social media sources because tweets can have different types of information and many attributes that the researcher can utilize (Kumar, 2013). These increase the validity of the research, along with internal and external validity.

Internal validity:

Babbie (2016) defines internal validity as "the possibility that the conclusions drawn from experimental results may not accurately reflect what went on in the experiment itself" (2016, p. 528). In order to ensure the internal validity of this research, tweets were collected from two different periods of time during the Covid-19 pandemic from the Panacea Lab. Additionally, several search sessions using the same keywords were conducted through the Twitter search

engine to ensure the retrieval of relevant data. Moreover, because Twitter contains many groups of tweets on subjects related to Covid-19, such as health, news, education, and business, a pilot study was conducted with a sample to determine information about the population of tweets relevant to the study of the users' behaviors in the early stage of the Covid-19 pandemic. Also, the final reporting of the results included the outcomes of sentiment analysis and LDA to ensure the internal validity.

In anticipation of potential issues of omission or commission that might be encountered in the implementation of the research design, such as node distribution, collection, and retrospective errors, which might weaken validity (Borgatti, 2013). The researcher collected all attributes such as links, videos, and photos linked with each tweet related to the Covid-19 pandemic collected from the Panacea Lab (Banda et al., 2020). Consequently, these threats did not create issues with the research design.

External validity:

According to Babbie (2016) determining external validity is important as "the possibility that conclusions drawn from experimental results may not be generalizable to the 'real' world" ((2016, p. 526). To ensure the external validity of the research the researcher chose a representative sample of user-provided data, 60 days before and 60 days after the declaration of the Covid-19 pandemic so the results of the study could be generalized it to the overall population and to provide a benchmark for further analyses following the pandemic. Also, because the Panacea Lap allowed the researcher to collect Big Data, the final results could be generalized to other social media platforms or other health crises, particularly on Twitter. Thus, the validity measures taken in this study ensure that the results can be generalized to the population under investigation as well as to other platforms and health emergencies.

Reliability:

Reliability refers to the stability and consistency of data collection and analysis processes (Gravetter, 2012), which are necessary for high quality results. Babbie (2016) defines reliability as "that quality of measurement method that suggests that the same data would have been collected each time in repeated observations of the same phenomenon" ((2016, p. 531). To avoid issues of instability and inconsistency in this research, data related to the Covid-19 pandemic were collected from Panacea Lab (Banda et al., 2020) were automatically instead of manually using various related keywords. As each tweet has many attributes, such as text, number of retweets, number of likes, etc., all different types of data were collected and stored on a personal computer, an external hard drive, and "UWM's Onedrive" cloud computing for backups.

To ensure the reliability of the research, two independent data analysts participated in the process using the same approach, and their results were compared with those of the research to determine degree of agreement among their coding results based on the Python and RStudio codes. Moreover, a pilot study was conducted to identify potential issues with the research design to ensure the validity and reliability of the study.

Summary:

The methodology chapter presents how to approach answering four research questions. The four research questions focused on comparing users' characteristics 60 days before and after the declaration of the Covid-19 pandemic. The tweets between January 11th, 2020, to May 10th, 2020, are the primary data collection for this study. The data cleansing processes were explained in this chapter. The sentiment analysis is used to answer the first research question, and the LDA model is used to answer the second research question. The inferential analysis tests were used to answer the third and fourth research questions

Chapter 4. Results:

This chapter presents the results of applying sentiment analysis and topic modeling to a comparison of Twitter users' posting behaviors during the 60 days before declaration of the Covid-19 pandemic, from January 11th, 2020, to March 10th, 2020, and their behaviors during the 60 days after the declaration, from March 11th, 2020, to May 10th, 2020

Descriptive Dataset:

As explained in the methodology chapter, on July 11th, 2021, all Covid-19 related tweets from January 11th, 2020 to May 10th, 2020 were collected from Panacea Lap (Banda et al., 2020) and hydrated. To ensure that all the texts were in English characters, only English tweets were extracted from the dataset by using the 'language' attribute and selecting the 'EN' language and "stri_enc_isascii" function in RStudio. The total number of English tweets was 21,655,283. Then the dataset was divided into two groups corresponding to the two time periods under investigation and referencing the research questions. The first group, representing the 60 days before the declaration of the Covid-19 pandemic, comprised 3,406,055 tweets. The second group, representing the 60 days after the declaration of the Covid-19 pandemic, comprised 18,249,229 tweets. Table 2 shows a brief description of the two groups.

Table 2. *Description of Collected the Covid-19 Tweets*

No. of	Group 1	Group 2	Total
1. Tweets	3,406,055	18,249,229	21,655,283
2. Unique users	733,219	2,714,759	3,447,978
3. Verified users	14,745	29,571	44,316
4. Non-verified users	718,474	2,685,188	3,403,662
5. Unique places	9,929	29,260	39,189
6. Unique hashtags	184,763	1,125,396	1,310,159
7. Retweets	14,693,952	57,514,036	72,207,988
8. Favorites	40,856,201	186,741,409	227,597,610

The tweets provided insights into the nature of users' interactions with regard to the Covid-19 pandemic. Altogether there were 3,447,978 unique users from 39,189 places around the world. Group 1 included 733,219 users from 9,929 locations, and group 2 included 2,714,759 users from 29,260 locations, indicating that the data came from various cultures, such as users from the U.S and U.K. Moreover, the users were classified into two categories, verified and non-verified users. Twitter's blue badge of verification lets people know that an account of public interest has been authenticated by the platform by meeting Twitter's criteria of verification. Tweets that Twitter can verify included those issued by governments, organizations, companies, influencers, etc. The proportion of verified users and places can enhance the quality of results. Table 3 shows a sample Covid-19 related tweet tweets after hydrating with all attributes collected from users. In this chapter, all users and links in tweet examples are indicated by stars for reasons of privacy and security.

Table 3. *Sample Dataset from Covid-19 Tweets with All Columns*

Attributes	Tweet Sample
coordinates	
created_at	Sat Jan 18 23:25:34 +0000 2020
hashtags	coronavirus China UK
media	
urls	https://*****
favorite_count	28
id	*******
in_reply_to_screen_name	*******
in_reply_to_status_id	*******
in_reply_to_user_id	*******
lang	en
place	
possibly_sensitive	FALSE
retweet_count	30
reweet_id	
retweet_screen_name	
source	<ahref=""https://mobile.twitter.com"" rel=""nofollow"">Twitter Web App
text	@****** @****** @****** The number of people already infected by the new #coronavirus emerging in #China is far greater than official figures. The #UK experts estimate a figure nearer 1,700. https://******
tweet_url	https://twitter.com/*****
user_created_at	Fri Jun 07 12:10:14 +0000 2019
user_screen_name	*****
user_default_profile_image	FALSE
user_description	#ChineseCommunistParty (#CCP) #pandemic #CCPLiePeopleDie. CCP imprison you at home. Seek the truth of #CCPVirus. Make CCP pay. #StandWithHongKong #EndTheCCP
user_favourites_count	5406
user_followers_count	6217
user_friends_count	2078
user_listed_count	30
user_location	United States

user_name	*******
user_screen_name.1	*******
user_statuses_count	13584
user_time_zone	
user_urls	https://www.youtube.com/**********
user_verified	FALSE

As shown, group 1 has considerably fewer tweets, unique users, unique places, unique hashtags, retweets, and favorites than group 2, indicating a sharp rise in interest after the pandemic had been declared. For example, the number of tweets after the announcement increased by 81.34%. Both groups were analyzed with the same techniques to obtain the 10 most repeated tweets for each group, which are shown in Table 4. Repeated tweets are duplicate tweets that are posted by a unique user, which is different from retweet. The main reason to find out the repeated tweets is to show how users copy and paste from others, which maybe bot. The users' names, IDs, and locations are not included for privacy reasons.

Table 4. *The 10 Most Repeated Tweets*

Group 1	Counts	Group 2	Counts
1. I believe that health is a human right, and #TaiwanCanHelp the @*** in preventing the spread of #COVID19.	1,760	1. We stand by Italy during these trying times. Share your Support for our Italian friends, They are our colleagues, friends and family. Cari amici, siamo con voi. #COVID19 #WeStandWithItaly	10,437
2. Reduce your risk from #coronavirus	1,603	2. . @***, I urge you to #cancelstudentdebt in the next #coronavirus package. A #StudentDebtStimulus will help the 45 million people with student debt and stimulate the economy when it is needed most.	1,424
3. Be Safe from #coronavirus infection	1,380	3. @*** @*** @*** @*** @*** @*** @***: To save as many lives as possible from #COVID19, we need the US and other world leaders to urgently fund vaccine development.	1,167

			Will you help fund @*** and stop #coronavirus?	
4.	Be Kind & support one another during #COVID19	508	4. A vaccine against #COVID19 could save millions of lives. @winstonpeters @***, will New Zealand step up and pledge new money so @*** can fund an urgent vaccine and help eliminate coronavirus across the globe?	1,096
5.	Be Smart & inform yourself about #coronavirus	505	5. Hey @*** @*** no country should have to choose between protecting their citizens from #COVID19 and paying their debts. I believe it's time to #CancelTheDebt and support a bold emergency response #GlobalCitizen	871
6.	@*** During the time of Lunar New Year, the Spring festival of China, a deadly outbreak of the coronavirus has emerged in the city of Wuhan. The question then arises,""What is the coronavirus? How dangerous is it?.""\n https://*****\n#Facts #Coronavirus #Health #Epidemi	215	6. We all need to work together to protect the world from #COVID19 and the May 4 pledging conference must be a success. Will @***states please step up and release much needed funds for response efforts to @*** @*** and @***? Together, we can beat this pandemic.	448
7.	Breaking #FoxNews Alert : First American coronavirus evacuees released from quarantine, pose 'no risk'	79	7. Be Safe from #coronavirus infection	372
8.	Breaking #FoxNews Alert : China sees sharp increase in coronavirus cases	78	8. Talk about fight against Coronavirus, advice people to wash their hands with sanitizer and stay away from crowded areas. Use the hashtag #FightCovid19	303
9.	Breaking #FoxNews Alert : Two planes carrying 350 American coronavirus evacuees from Wuhan, China, land at Travis Air Force Base in California	77	9. . @*** @***, reports around the country show that people of color are dying of COVID-19 at highly disproportionate rates. Please ensure states collect and report comprehensive data on infection, death and testing rates by race so that we properly protect the most vulnerable	291
10.	Breaking #FoxNews Alert : Coronavirus death in Philippines said to be first outside China	77	10. CHINA must be dragged into International Court and stripped of its VETO power in the UN\n 'Crime against humanity'\nCOVID-19 is a Chinese Virus. \n\nCopy and paste.	280

Hashtags can be an important factor that increases the appearance of tweets and interactions among users on Twitter. Some frequently used hashtags appeared in both groups. Table 5 presents the 10 most frequently used hashtags in tweets.

Table 5. *The Top 10 Appearances of Hashtag*

	Group 1	Count		Group 2	Count
1.	**#Coronavirus**	531,992	**1.**	**#COVID19**	1,950,555
2.	**#COVID19**	161,139	**2.**	**#coronavirus**	1,264,925
3.	**#China**	51,748	3.	#Lockdown	80,946
4.	#Wuhan	36,055	4.	#StayHome	75,227
5.	#Virus	25,351	5.	#Pandemic	61,675
6.	#Flu	19,904	**6.**	**#Trump**	46,618
7.	#Sars	17,574	7.	#SocialDistancing	44,335
8.	**#Trump**	16,075	**8.**	**#China**	41,680
9.	**#News**	14,952	9.	#IndiaFightsCorona	37,619
10.	#Iran	10,062	**10.**	**#News**	37,019

Note: Repeated hashtags are on bold

Although, due to the nature of tweets, it is not possible to know how often particular tweet has been viewed, the number of retweets and favorite ratings it has received provide some indication of how many users have seen it. Users can use retweet function to instantly share information to a large number of audiences using direct or indirect lists of followers. As a method by which users can join a conversation, the retweet function can be used to engage with and validate others while avoiding the large issues such as authorship, attribution, and communicative fidelity (Boyd et al., 2010). Table 6 presents the 10 most frequently retweeted tweets have highest number of retweets, and Table 7 shows the 10 tweets most frequently rated as favorites.

Table 6. *The 10 Tweets with Highest Number of Retweets*

Group 1	No. of Retweets	Group 2	No. of Retweets
1. Coronavirus has crossed the line for Italians \nhttps://*****	200,224	1. COVID-19 helping people realise that some meetings can be emails.	194,052
2. 900 people get Coronavirus and the whole world wants to wear surgical mask, 30 million people have AIDS but still nobody wants to wear a condom https://*****	156,207	2. What coronavirus symptoms look like, day by day https://*****	182,866
3. PSA!!! facial mask is not effective against wuhan virus https://*****	148,090	3. A message from me and my dad, @***. \n\n#coronavirus #DontBeASpreader https://*****	181,017
4. Scientists: you should wash your hands because of Coronavirus.\n\nPeople: I'm gonna stop flying, hoard masks, work from home & totally rearrange my life.\n\nAlso Scientists: the #ClimateCrisis will kill millions - we must use clean power & change how we get to work.\n\nPeople: No way.	130,188	4. This is HARD AS F***. https://*****	133,483
5. the coronavirus outbreak in Italy https://*****	116,530	5. maybe ur not scared of coronavirus but u should be scared about passing it on to ppl you know who are at risk. thats not just old people - its ppl with asthma, ppl who smoke, ppl who have underlying health problems (aka a lot more people than you think)	131,574
6. Coronavirus really closing everything down in the country, but my job	83,219	6. If you believe the world will overcome covid-19 retweet. https://*****	117,405
7. coronavirus has been racialised as a ""chinese"" illness and for this reason chinese people, regardless of their proximity to wuhan are being treated like carriers of the virus. similar to	79,777	7. After COVID-19 is over, I better NEVER hear anyone trash ""low end"" workers again. Those people at the grocery store, the Dollar General workers, those fast food workers, the Walmart employees, those people you	111,256

	how ebola was subtly touted as an ""african"" illness. https://*****		didn't even think deserved to have a wage to survive on?	
8.	BREAKING: A D.C. priest has Coronavirus. He offered communion and shook hands with more than 500 worshippers last week and on February 24th. All worshippers who visited the Christ Church in Georgetown must self-quarantine. Church is cancelled for the first time since the 1800's	77,845	8. coronavirus this coronavirus that.... can we talk about me for a second??	98,297
9.	JOE BIDEN LITERALLY SAID HE'LL VETO MEDICARE FOR ALL.\n\nDURING THE CORONAVIRUS OUTBREAK.\n\nON TV.\n\nTHIS MAN IS NOT ELECTABLE IN NOVEMBER.	74,863	9. Halting funding for the World Health Organization during a world health crisis is as dangerous as it sounds. Their work is slowing the spread of COVID-19 and if that work is stopped no other organization can replace them. The world needs @*** now more than ever.	91,885
10.	Vietnam was the first country to fully contain SARS and COVID-19 (with no deaths) and developed a quick-test kit in one month that the WHO says should have taken four years. WHO is now consulting with Vietnam to get help for the global crisis. Of course this isn't in our news.	73,198	10. WHAT https://*****	88,304

Table 7. *The 10 Tweets with Highest Number of Favorite Ratings*

Group 1	No. of Favorite	Group 2	No. of Favorite
1. Coronavirus has crossed the line for Italians \nhttps://*****	670,134	1. COVID-19 helping people realise that some meetings can be email	773,654
2. 900 people get Coronavirus and the whole world wants to wear surgical mask, 30 million people have AIDS but still nobody wants to wear a condom https://*****	483,449	2. This is HARD AS F***. https://*****	674,819

3.	Scientists: you should wash your hands because of Coronavirus.\n\nPeople: I'm gonna stop flying, hoard masks, work from home & totally rearrange my life.\n\nAlso Scientists: the #ClimateCrisis will kill millions - we must use clean power & change how we get to work.\n\nPeople: No way.	443,573	3.	maybe ur not scared of coronavirus but u should be scared about passing it on to ppl you know who are at risk. thats not just old people - its ppl with asthma, ppl who smoke, ppl who have underlying health problems (aka a lot more people than you think)	541,202
4.	the coronavirus outbreak in Italy https://*****	425,209	4.	A message from me and my dad, @***. \n\n#coronavirus #DontBeASpreader https://*****	482,248
5.	I've heard that coronavirus is going to cause a massive shortage of books, which will be essential when we're all stuck at home, so it's very important for everyone to rush out and start panic-buying novels. Thank you.	402,797	5.	WHAT https://*****	481,058
6.	Coronavirus really closing everything down in the country, but my job	345,733	6.	After COVID-19 is over, I better NEVER hear anyone trash ""low end"" workers again. Those people at the grocery store, the Dollar General workers, those fast food workers, the Walmart employees, those people you didn't even think deserved to have a wage to survive on?	456,458
7.	JOE BIDEN LITERALLY SAID HE'LL VETO MEDICARE FOR ALL.\n\nDURING THE CORONAVIRUS OUTBREAK.\n\nON TV.\n\nTHIS MAN IS NOT ELECTABLE IN NOVEMBER.	324,296	7.	If your job is so ""essential"" that you can't get off for a killer global pandemic, you deserve $15 an hour and a union.	444,438
8.	Vietnam was the first country to fully contain SARS and COVID-19 (with no deaths) and developed a quick-test kit in one month that the WHO says should have taken four years. WHO is now consulting with Vietnam to get help for the global crisis. Of course this isn't in our news.	277,948	8.	jared leto having no idea about coronavirus because he was on a meditative retreat in the desert is the most jared leto thing in the world https://*****	390,699

9.	BREAKING: A D.C. priest has Coronavirus. He offered communion and shook hands with more than 500 worshippers last week and on February 24th. All worshippers who visited the Christ Church in Georgetown must self-quarantine. Church is cancelled for the first time since the 1800's	265,764	9. Am I the only one who goes to bed every night convinced that my mild throat scratch is the beginnings of coronavirus only to wake up with a wave of huge relief that it's gone?	387,190
10.	coronavirus has been racialised as a ""chinese"" illness and for this reason chinese people, regardless of their proximity to wuhan are being treated like carriers of the virus. similar to how ebola was subtly touted as an ""african"" illness. https://*****	256,450	10. coronavirus this coronavirus that.... can we talk about me for a second??	376,492

Some tweets had no inherent meaning but only external links for reference or marketing purposes. For instance, this tweet 'WHAT https://*****' has an external link to news or some other site. We removed the links and users for purposes of privacy and security. In some tweets culturally specific slang was used, making them meaningless for most readers.

Findings Related to The Sentiment Characteristics:

1. What are the key sentiment characteristics of tweets related to Covid-19 posted during the 60 days before and the 60 days after declaration of the Covid-19 pandemic?

This research question addresses the sentiment and emotional characteristics of tweets appearing during the two target periods of the Covid-19 pandemic.

Sentiment Analysis:

To develop a sentiment classifier model, the cleansed tweets (from which all punctuation, numbers, lowercasing, stop words, white space, and stemming were removed as shown in Figure 3 in the methodology chapter) were scored and classified by polarity (positive, neutral, or negative). To achieve this, the VADER sentiment analysis tools in Python were used. Each

tweet's polarity was labeled as positive, negative, or neutral. It was labeled as positive if its polarity was greater than 0.05, as neutral if its polarity was less than 0.05 and greater than -0.05, and as negative if its polarity was less than -0.05. The number of tweets in each category was recorded, and their percentages of the total number of tweets for each group were calculated. Table 8 summarizes the results of the sentiment analysis for both groups.

Table 8. *Sentiment Characteristics for Both Groups*

	Group 1		**Group 2**	
Sentiment Analysis	**No. of Tweets**	**Percentage**	**No. of Tweets**	**Percentage**
Positive Tweets	928,961	27.3%	6,982,729	28.7%
Neutral Tweets	1,114,181	32.7%	5,244,152	33%
Negative Tweets	1,362,913	40%	6,022,348	38.3%
Total	**3,406,055**	**100%**	**18,249,229**	**100%**

Emotional Analysis:

To conduct emotional analysis of Covid-19 related tweets during the first 60 days before and the 60 days after the pandemic was declared, the cleansed tweets were scored and classified into 10 categories using 'syuzhet' package in RStudio. These categories included eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). Negative tweets constituted the highest percentage for users in group 1, and positive tweets constituted the highest percentage for users' emotion in group 2. Although Python was employed for the sentiment analysis, the 'syuzhet' package in RStudio also employed for determining positive and negative sentiment characteristics. Table 9 and Figure 5 show the percentages of users' emotions and sentiment for the two groups.

Table 9. *Emotional Analysis for Both Groups*

Emotion	Group 1		Group 2	
	No. of Tweets	Percentage	No. of Tweets	Percentage
Anger	161,461	4.74%	839,986	4.60%
Anticipation	391,198	11.49%	2,093,038	11.47%
Disgust	134,939	3.96%	699,148	3.83%
Fear	349,481	10.26%	1,401,949	7.68%
Joy	114,673	3.37%	791,015	4.33%
Sadness	221,059	6.49%	1,067,425	5.85%
Surprise	187,331	5.50%	992,612	5.44%
Trust	366,203	10.75%	2,429,731	13.31%
Negative	882,087	25.90%	3,620,263	19.84%
Positive	597,623	17.55%	4,314,062	23.64%
Total	**3,406,055**	**100.00%**	**18,249,229**	**100.00%**

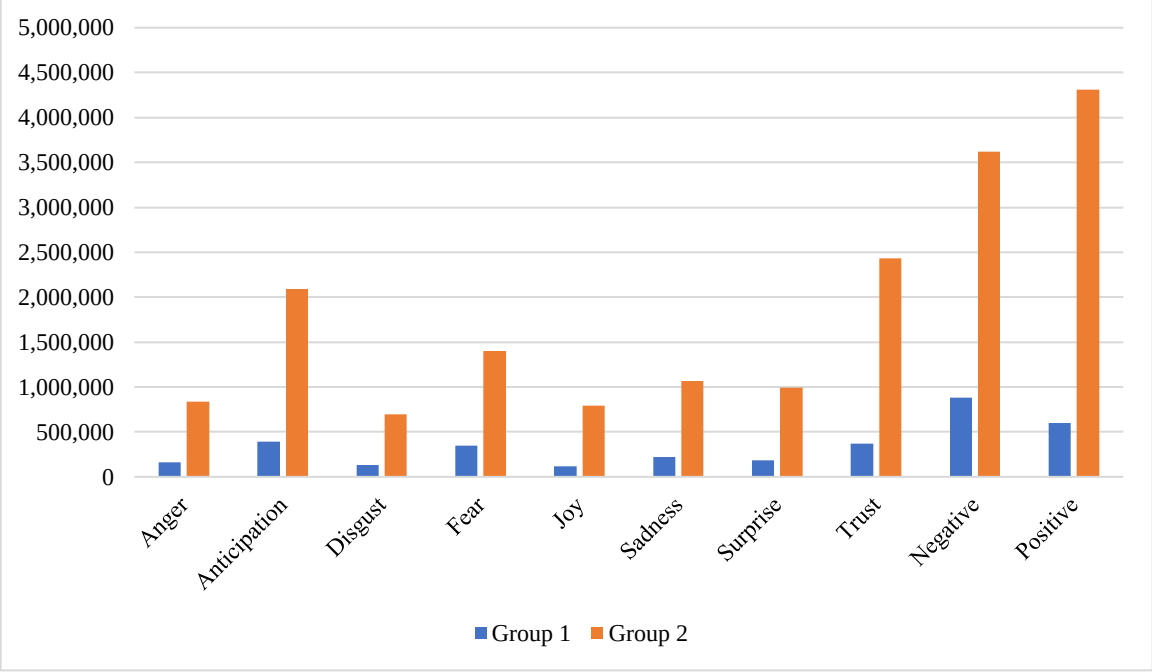

Figure 5. Emotional Analysis for Both Groups

Findings Related to The Discussion Topics:

2. What are the key discussion topics emerged from tweets related to Covid-19 during the 60 days before and the 60 days after the declaration the Covid-19 pandemic?

One of the prime goals of this research was to discover the topics of discussions about the Covid-19 pandemic on Twitter 60 days before and 60 days the declaration of the pandemic. The LDA model was used to determine the topics that emerged from the users' tweets.

Model training and evaluation processes were used for this analysis. The cleansed tweets underwent a preparation process for the LDA model. The input datasets for each group consisted of tweets generated by users. Each tweet was processed by text preparation procedures including tokenization, stop-word removal, lemmatization, punctuation removal, lowercase correction, stemming, number removal, users' id removal, url removal, and extra whitespace removal (see Figure 3 in the methodology chapter).

LDA uses a machine learning technique. Among the different LDA models in Python, the LDA model using the sklearn tool was used to discover the topics discussed. Table 10 shows the results of LDA model analysis with topic weights for group 1 and Table 11 for group 2. After developing suitable LDA models, the discovered topics in both groups were labeled manually based on the most frequently used terms in the topic-term distributions. Labelling topics made it possible to interpret the corpus regarding which concepts were prevalent (Saeidi et al., 2015). Sievert and Shirley (2014) suggest that topics can be interpreted by examining a ranked list of the most probable terms in that topic. Thus, the topic interpretation is subjective.

Table 10 *LDA Result for Group 1*

Topic 1 Health	**Topic 1 weights**	**Percentage**	**Topic 2 Wuhan**	**Topic 2 weights**	**Percentage**
Health	137234	24.6%	Wuhan	210692.5	39.5%
Trump	71242.8	12.8%	China	111954.8	21.0%
China	57661.5	10.3%	Chinese	56637.3	10.6%
Emergency	52812.3	9.5%	Doctor	34107.3	6.4%
Public	49122	8.8%	Hospital	30082.9	5.6%
Sars	43222.7	7.7%	Outbreak	26025.7	4.9%
Flu	42827.5	7.7%	Video	19619.1	3.7%
Get	38296.1	6.9%	City	14969.2	2.8%
Official	32906.6	5.9%	Animal	14963.2	2.8%
Human	32464.7	5.8%	People	13855.7	2.6%
Topic 3 News	**Topic 3 weights**	**Percentage**	**Topic 4 Market**	**Topic 4 weights**	**Percentage**
News	113716.9	23.0%	Market	63236.2	17.1%
Update	65928.8	13.3%	Stock	52621.2	14.2%
Confirmed	64125.5	13.0%	Spread	45150	12.2%
Patient	45069.5	9.1%	Vaccine	43987.8	11.9%
Live	43320	8.8%	Global	39298.4	10.6%
Uk	38812.9	7.8%	Fear	35846.7	9.7%
School	32146.2	6.5%	Economy	24477.4	6.6%
Health	31169.8	6.3%	Disease	24146.6	6.5%
Rate	31105.2	6.3%	Cure	20985.9	5.7%
Tested	29209.4	5.9%	Test	20000.5	5.4%
Topic 5 China	**Topic 5 weights**	**Percentage**	**Topic 6 Flight**	**Topic 6 weights**	**Percentage**
China	181916.7	23.9%	China	45863.8	18.1%
Death	178498.2	23.5%	Flight	33977.2	13.4%
Report	67628.4	8.9%	Impact	27007.1	10.6%

Number	59799.7	7.9%	Hit	23505.9	9.3%
Toll	55601.1	7.3%	County	23402.4	9.2%
Country	49449.9	6.5%	House	22531.2	8.9%
Iran	46118.3	6.1%	Business	20477.5	8.1%
Confirmed	43366.3	5.7%	Airline	20245.1	8.0%
State	39871.4	5.2%	Amid	19791.9	7.8%
Rise	37563.1	4.9%	Apple	16975.3	6.7%
Topic 7 Countries	**Topic 7 weights**	**Percentage**	**Topic 8 Cruise**	**Topic 8 weights**	**Percentage**
China	62238.4	14.0%	Cruise	76297.6	20.1%
Day	61623.4	13.9%	Ship	68344.1	18.0%
Italy	54889.2	12.4%	Japan	42989.1	11.3%
South	53162.5	12.0%	News	33902.6	8.9%
Korea	50343.2	11.4%	Passenger	33581.1	8.8%
Wuhan	48126.1	10.9%	Quarantined	31337.8	8.2%
Quarantine	47799.1	10.8%	Test	30238.3	8.0%
Spread	22251.1	5.0%	Quarantine	21447.6	5.6%
Chinese	22098	5.0%	American	20925.3	5.5%
Hospital	20645	4.7%	People	20913.5	5.5%
Topic 9 Hygiene	**Topic 9 weights**	**Percentage**	**Topic 10 People**	**Topic 10 weights**	**Percentage**
Hand	49125.4	17.8%	People	136318.8	29.9%
Latest	38673.2	14.0%	Trump	75010.8	16.5%
Daily	34942.2	12.7%	Flu	44816.9	9.8%
Fear	25303.1	9.2%	Going	41276.3	9.1%
Event	25068.6	9.1%	Thing	35008.2	7.7%
Concern	23524.3	8.5%	Time	28206.3	6.2%
Cancel	22779.3	8.3%	Die	24669.5	5.4%
Mask	19665.5	7.1%	Pneumonia	24367.3	5.3%
Amid	19623.2	7.1%	Kill	23312.8	5.1%
Stay	17156	6.2%	Sick	22951.1	5.0%

Labeling Ten topics based on the LDA outputs is a fundamental aspect to manually label the Ten terms under One topic. Topic 1 as (Health) discusses the early stage of the Covid-19 when people did not know what the Covid-19 exactly is. Topic 2 as (Wuhan) contains keywords described where the Covid-19 started. Topic 3 as (News) includes reporting the Covid-19 cases and rate. Topic 4 (Market) described how the Covid-19 affected on economy. Topic 5 (China) reports the Covid-19 cases and deaths how increased. Topic 6 (Flight) explains how the airlines got impact. Topic 7 (Countries) shows different countries started quarantine. Topic 8 (Cruise) described the cruise stuck in the ocean. Topic 9 (Hygiene) talks about how to protect people from the Covid-19. Topic 10 (People) describes the individuals' bad situation.

Table 11 *LDA Result for Group 2*

Topic 1 Trump	**Topic 1 weights**	**Percentage**	**Topic 2 Lockdown**	**Topic 2 weights**	**Percentage**
Trump	967850.8	29.9%	Lockdown	585037.2	18.1%
Stay	371909.7	11.5%	Country	486288.8	15.1%
Business	333179	10.3%	Good	400848.2	12.4%
American	310064.2	9.6%	Testing	349869.4	10.8%
President	260178.2	8.0%	Crisis	334906.2	10.4%
Call	224651.8	7.0%	Year	329074.1	10.2%
Safe	206100.1	6.4%	India	259028.1	8.0%
Hope	200098.3	6.2%	School	175778.7	5.4%
Vaccine	188438	5.8%	Hand	162831	5.0%
Post	169920.3	5.3%	Flu	142961.1	4.4%
Topic 3 Death	**Topic 3 weights**	**Percentage**	**Topic 4 Update**	**Topic 4 weights**	**Percentage**
Death	737051.3	25.0%	Update	369038.1	15.4%
Time	554342.9	18.8%	Thing	284206.9	11.9%

Read	285534.8	9.7%	Live	271705.1	11.3%
Uk	264095.8	9.0%	Death	253127.1	10.6%
Number	208102.7	7.1%	Total	240536.2	10.0%
Free	204955.4	7.0%	Report	233037.7	9.7%
Rate	176536.6	6.0%	Data	196166.5	8.2%
Mask	175502.5	6.0%	Confirmed	195370.5	8.2%
Better	172347.9	5.8%	April	179891.5	7.5%
Care	169987	5.8%	Dr	173465.5	7.2%
Topic 5 Pandemic	**Topic 5 weights**	**Percentage**	**Topic 6 People**	**Topic 6 weights**	**Percentage**
Pandemic	1193009	37.4%	People	1342340	37.7%
Health	519058.2	16.3%	Day	607146.2	17.1%
Public	251582.1	7.9%	Week	359775	10.1%
Amid	236561.7	7.4%	Today	229967.5	6.5%
Global	203956.4	6.4%	Daily	210494.7	5.9%
Risk	193666.4	6.1%	Latest	195091.6	5.5%
Long	155418.3	4.9%	Died	173357.8	4.9%
Change	153749.7	4.8%	Die	166413.6	4.7%
Start	144302.4	4.5%	Real	162400.1	4.6%
System	134662.9	4.2%	Travel	113101.3	3.2%
Topic 7 Help	**Topic 7 weights**	**Percentage**	**Topic 8 Life**	**Topic 8 weights**	**Percentage**
Help	600248.7	24.1%	China	502188.1	20.3%
Spread	372797.9	15.0%	Work	405865.9	16.4%
Support	317090.8	12.7%	Life	279989.4	11.3%
Family	267370.6	10.7%	Social	258243.8	10.5%
Money	180677.1	7.3%	Job	202693.8	8.2%
Best	161676.5	6.5%	Medium	181250.2	7.3%
Share	160395.1	6.4%	God	168044.5	6.8%
Human	147916.3	5.9%	Man	165280.2	6.7%
Pm	143618.3	5.8%	Distancing	153478.7	6.2%

Friend	136529.1	5.5%	Article	151550.5	6.1%
Topic 9 Hospital	**Topic 9 weights**	**Percentage**	**Topic 10 News**	**Topic 10 weights**	**Percentage**
State	560114.1	19.3%	News	697420.2	26.1%
Test	512043.6	17.7%	Government	456755	17.1%
Patient	389478.1	13.4%	Response	324281.3	12.1%
Hospital	334935.5	11.6%	Fight	276243.4	10.3%
Going	275752.3	9.5%	Face	192427.8	7.2%
Doctor	242208	8.4%	House	176905.1	6.6%
Staff	161464.1	5.6%	Medical	151236.2	5.7%
Child	147873	5.1%	Leader	151027.5	5.7%
Person	140254.8	4.8%	Woman	121902	4.6%
Issue	132757.2	4.6%	White	121591.1	4.6%

As shown in the Table 11, users discussed different subjects after declaration of the Covid-19 pandemic. Topic 1 (Trump) contains terms described President Trump on news conference about the Covid-19. Topic 2 (Lockdown) is the beginning of the government's restrictions. Topic 3 (Death) talks about wearing masks can reduce the death from the Covid-19. Topic 4 (Update) discusses the number of cases in April. Topic 5 (Pandemic) declares the Covid-19 pandemic. Topic 6 (People) discusses the number of cases and deaths daily. Topic 7 (Help) converse about seeking help to support family. Topic 8 (Life) describes how social distancing change the life. Topic 9 (Hospital) talks about how health system faced an issue due to number of patients. Topic 10 (News) reports how governments response and fight the Covid-19.

LDA Model Evaluation:

To ensure the quality of results, it is essential to evaluate the LDA model. The pyLDAvis imported Python package, which is a port of the R package created by Sievert and Shirley (Mabey, 2015), is designed to assist users to interpret topic models which have been

employed for text analysis. This package generates information and results based on LDA topic models and forms web-based visualization plots. The main visualization tool of this package plots topic circles (bubbles) and word frequencies interactively for in-depth evaluation. The range of LDA topics was tuned to reach a set of non-overlapping clusters that had sufficient distances among them. For instance, given the K parameter as ten, seven of the ten resulting topics represented as circles in the inter-topic distance map overlapped with each other as shown in Figure 6. The main reason is that pyLDAvis uses mds default parameters to visualize the topics using Principal Coordinate Analysis (PCA) with a distance matrix created using the Jensen-Shannon divergence on the topic-term distributions.

Figure 6. PCA of Evaluation of the LDA Model

In this study, pyLDAvis was used to evaluate the LDA model to find the best number of topics as the K parameter. The evaluation model indicated that ten topics appeared in the Covid-19 tweets for both groups 1 and 2. The results of applying the LDA model showed that topics were interpretable based on their terms. The ten topics emerged from Covid-19 on Twitter during the first 60 days before declaring the Covid-19 pandemic were (*Health, Wuhan, News, Market, China, Flight, Countries, Cruise, Hygiene, and People*). The ten topics that emerged during the 60 days after the declaration of the Covid-19 pandemic were *Trump, Lockdown, Death, Update, Pandemic, People, Help, Life, Hospital, and News*). A different multidimensional scaling (mds) parameter was used to avoid overlapping topics, Metric Multidimensional Scaling (mmds). The maps of inter-topic distances are shown in Figure 7 for group 1 and Figure 8 for group 2. These maps show there was no overlapping among topics. The pyLDAvis using different methods thus offers the best visualization to show the topics-terms distribution.

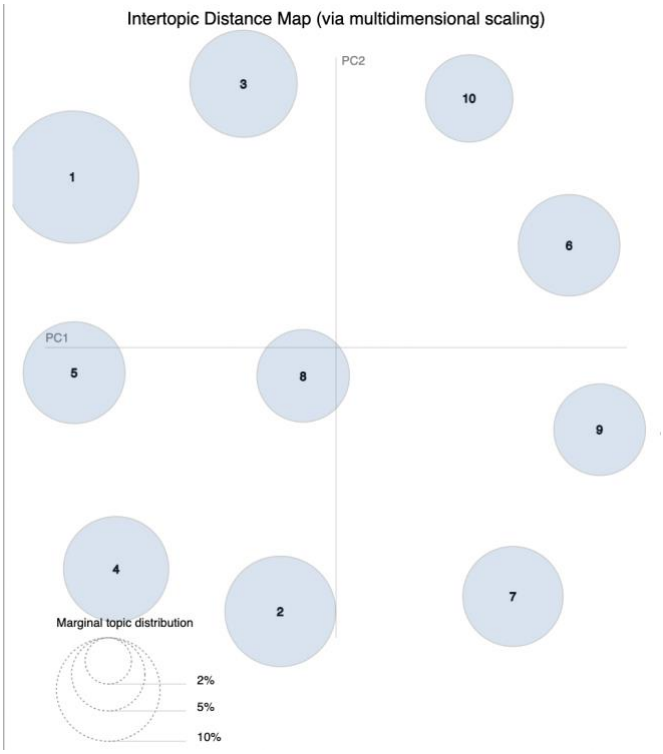

Figure 7. Evaluation of the LDA Model via MMDS Method for Group 1

Figure 8. Evaluation of the LDA Model via MMDS Method for Group 2

The pyLDAvis has two main parts, which are the topics and the 30 most relevant terms. The topics are shown in Figures 7 and 8 as circles on a two-dimensional plane, the centers of which were determined by computing the Jensen–Shannon divergence between topics, and then by using multidimensional scaling to project the inter-topic distances onto two dimensions. Each topic's overall prevalence is encoded using the zones of the circles. The top 30 terms for Topic 1 are shown as an example in Figure 9, which depicts a horizontal bar chart in which the bars represent the individual terms that are the most useful for interpreting the selected topic on the left. Each pair of conjoined bars represent both the corpus-wide frequency of a given term as well as the topic-specific frequency of the term.

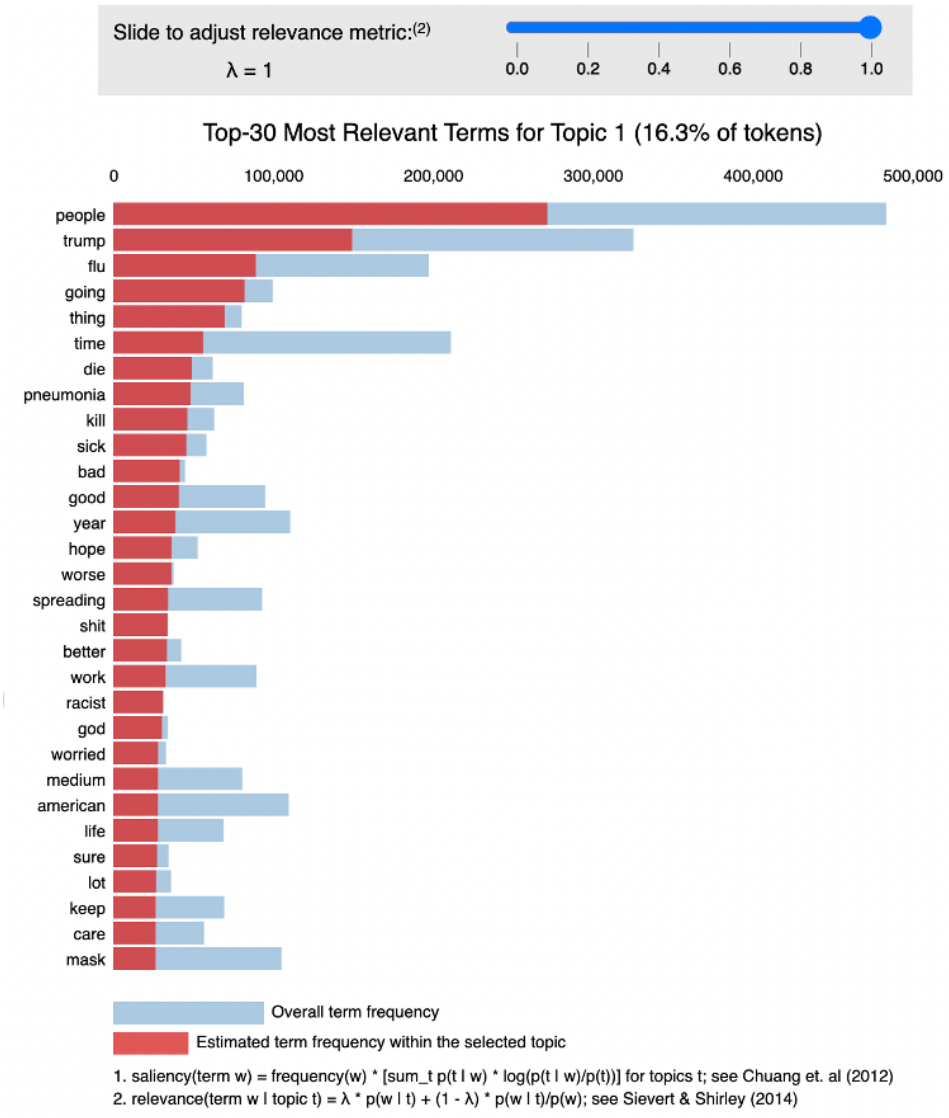

Figure 9. Top-30 Most Relevant Terms for Topic 1 in Group 1

Findings Related to The Relationships Between Sentiment Characteristics and Discussion Topics:

3. What are the relationships between sentiment characteristics and discussion topics in Covid-19 related tweets during the 60 days before and the 60 days after the declaration of the Covid-19 pandemic?

Research question three addresses the relationship between the results of RQ1 (sentiment characteristics) and RQ2 (discussion topics). To answer this this research question, sentiment and emotional analysis were applied to analyze each tweet to reveal sentiment characteristics, and LDA was applied analyze each tweet to reveal discussion topics. Table 12 shows the total tweets based on discussion topics and sentiment characteristics for group 1 and Table 13 for group 2. Sankey diagram can help locate the most important contributions to a flow between sentiment characteristics and discussion topics. Figures 9 and 10 illustrate the relationships between sentiment characteristics and discussion topics for group 1 and 2.

Table 12 *The Distribution of Sentiment Characteristics in Discussion Topics for Group 1*

Topics	Sentiment							Total %
	Negative	%	Neutral	%	Positive	%	Total	
1. Health	137,953	38.63%	125,506	35.14%	93,670	26.23%	**357,129**	**10.49%**
2. Wuhan	127,755	41.20%	101,207	32.64%	81,129	26.16%	**310,091**	**9.10%**
3. News	92,718	29.06%	151,262	47.40%	75,117	23.54%	**319,097**	**9.37%**
4. Market	147,861	38.76%	118,913	31.17%	114,719	30.07%	**381,493**	**11.20%**
5. China	141,268	45.41%	116,384	37.41%	53,467	17.19%	**311,119**	**9.13%**
6. Flight	92,000	37.86%	100,711	41.45%	50,270	20.69%	**242,981**	**7.13%**
7. Countries	90,074	37.68%	94,084	39.35%	54,922	22.97%	**239,080**	**7.02%**
8. Cruise	84,727	39.42%	72,889	33.91%	57,328	26.67%	**214,944**	**6.31%**
9. Hygiene	119,005	30.76%	112,980	29.20%	154,903	40.04%	**386,888**	**11.36%**

10.People	329,552	51.23%	120,245	18.69%	193,436	30.07%	**643,233**	**18.88%**
Total	**1,362,913**	**40%**	**1,114,181**	**33%**	**928,961**	**27%**	**3,406,055**	**100.00%**

Note: % = Percentage

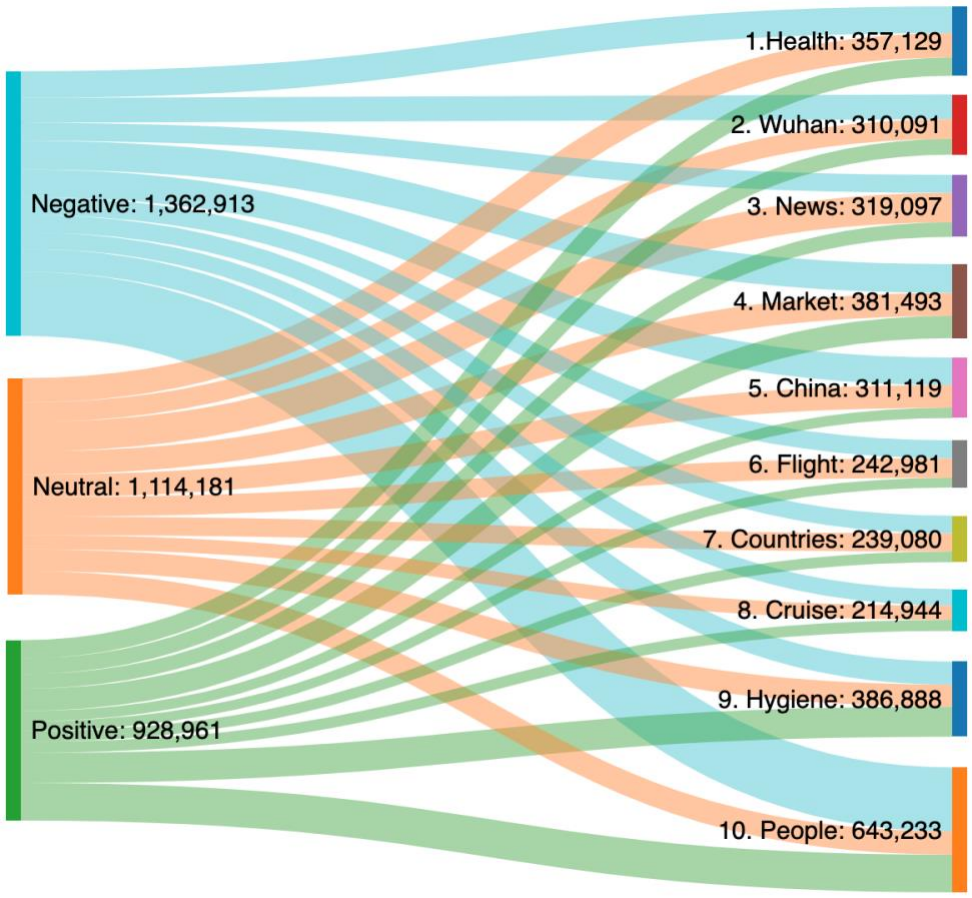

Figure 10. Sankey Diagram of Sentiment Characteristics in Discussion Topics for Group 1

Table 13 *The Distribution of Sentiment Characteristics in Discussion Topics for Group 2*

Topics	Sentiment							Total %
	Negative	**%**	**Neutral**	**%**	**Positive**	**%**	**Total**	
1. Trump	1,667,478	29.19%	2,087,135	36.54%	1,957,252	34.27%	**5,711,865**	**31.30%**
2. Lockdown	626,097	35.72%	427,799	24.41%	698,748	39.87%	**1,752,644**	**9.60%**
3. Death	613,639	38.56%	358,212	22.51%	619,359	38.92%	**1,591,210**	**8.72%**
4. Update	374,326	28.40%	502,149	38.09%	441,698	33.51%	**1,318,173**	**7.22%**
5. Pandemic	658,425	36.98%	440,469	24.74%	681,781	38.29%	**1,780,675**	**9.76%**

6. People	519,061	42.70%	273,002	22.46%	423,665	34.85%	**1,215,728**	**6.66%**
7. Help	250,838	20.71%	199,758	16.49%	760,711	62.80%	**1,211,307**	**6.64%**
8. Life	418,715	34.48%	291,594	24.01%	504,055	41.51%	**1,214,364**	**6.65%**
9. Hospital	449,260	33.43%	358,766	26.70%	535,733	39.87%	**1,343,759**	**7.36%**
10. News	444,509	40.06%	305,268	27.51%	359,727	32.42%	**1,109,504**	**6.08%**
Total	6,022,348	33%	5,244,152	28.74%	6,982,729	38.26%	**18,249,229**	**100%**

Note: % = Percentage

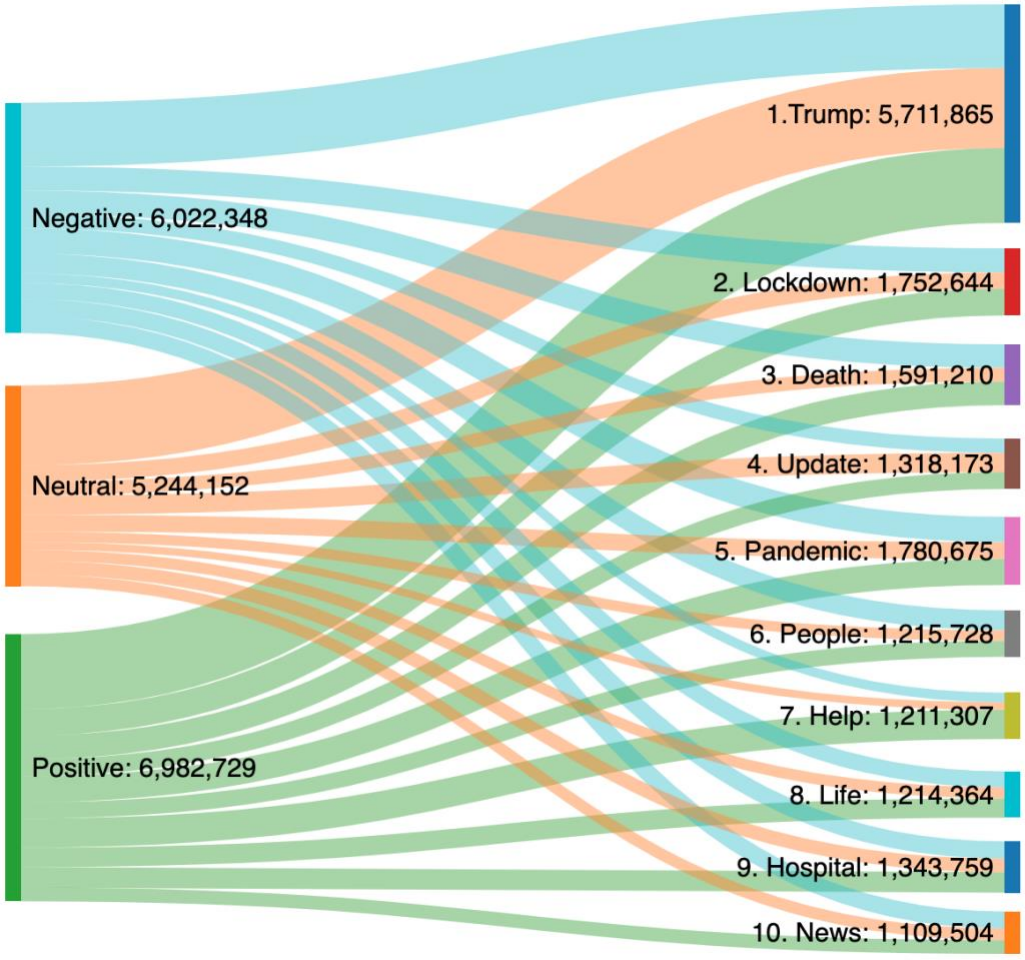

Figure 11. Sankey Diagram of Sentiment Characteristics in Discussion Topics for Group 2

We retrieved a random sample of tweets from each topic with different sentiment characteristics as shown in table 14 for group 1 and table 15 for group 2.

Table 14 *Sample Tweet from Each Topic and Sentiment for Group 1*

	Topic	Tweet	Sentiment
1.	***Health***	Great story on a great medical center where I trained. @*** front and center caring for those with coronavirus. Great foresight by their community and leading the country in keeping us all safe. https://*****	Positive
		Differences between #COVID19 and influenza. The World Health Organization says that so far from data in China it looks like adults spread COVID-19 to children (somewhat the opposite for flu). That and more: https://***** https://*****	Neutral
		Why does the World Health Organization\nrefuse to declare an emergency\ndue to CORONAVIRUS?\n\nWhy? They know it's a FAKE.\nIts a cyber attack derived by\nWestern Powers out to punish\nChina.\n\nFear is hurting China as it\nloses sales of its products\ndue to world fear.\n\n#rogerbezanis https://*****	Negative
2.	***Wuhan***	Prayers for Wuhan people in China, where a lockdown is for everyone there. We ask you for prayers of love. Let's not abandon them. Let them know we care. Us the Indigenous people of the world and others ..share our compassionate heart. We care and we love-this is human kindness.	Positive
		I pray for these folks who cannot stop what China is doing to them! Fight Wuhan Fight and all the other cities : FIGHT!\n\nhttps://*****	Neutral
		Wuhan corpse truck full of dead bodies\nThe epidemic in Wuhan has continued to expand, with several dead cities in the epidemic area. Corpse trucks are running all over the city. Even ambulances have been changed to corpse trucks.\nNO TREATMENT - just collect the dead! https://*****	Negative
3.	***News***	My sister is one of the many courageous nurses caring for patients with the #coronavirus (#COVID19) in Washington. She has had two positive patients so far. The good news is that she said she's unafraid but I will be more reassured when nurses have better access to testing. https://*****	Positive
		Some sources say the coronavirus has been confirmed in Virginia and some say it hasn't which is really helpful to a public that's prone to panicking	Neutral
		Caught By Surprise: Lack Of Coronavirus Testing Gives U.S. Highest Reported Mortality Rate https://***** #SmartNews	Negative
4.	***Market***	Joe Biden's Super Tuesday success could help calm coronavirus-rocked market - https://*****\nWant to know more on Joe Biden's Super Tuesday success could help calm coronavirus...	Positive
		The slowing cases of #coronavirus\nhave given #oil prices a boost, reaching a new high since January 29. However, despite the supply shocks from #Venezuela and	Neutral

	#Libya, experts warn that the worst of the economic impact may not be over.\n\nhttps://***** https://*****	
	TREASURIES-U.S. 10-year yield hits new low as coronavirus impacts... https://*****	Negative
5. China	The new coronavirus has also begun to spread in the United States. It is good to see Trump attach great importance. At the same time, he hopes that the United States can learn from China's resolute and rapid response to make the virus preventable and controllable. https://*****	Positive
	@*** Today's Wuhan Coronavirus report out of China, usually around 5 p.m. ET (delayed yesterday), should be particularly important for direction.	Neutral
	#COVID19: Death toll in China nears 2600-mark. Infection cases in South Korea surge past 800 with eight deaths while in Italy, sixth death reported. Wall Street falls on pandemic fears. Health Minister Dr Harsh Vardhan holds review meeting https://*****	Negative
6. Flight	@*** I wouldn't offer any special flights now, this coronavirus is getting out of hand. But once this is over, I'm pretty sure that . @*** would definitely like to enjoy the white sand you speak of. New direct flights from Ho Chi Minh City to .@*** via Osaka. https://*****	Positive
	@*** That I totally agree, and China has been applying this rule since Wuhan shutdown. But this is not same as ""blocking every plane and everyone""	Neutral
	Trudeau is Canada's biggest threat\n\nThe horribly incompetent and irresponsible way his team handled Coronavirus \n\nHe refused to shut down flights from infected countries. Now US is shutting down Canadians \n\nU.S. border refusals to include Canadians\n\nhttps://*****	Negative
7. Countries	Italy in these days is a protected area for coronavirus and these scenes do not take place, but we will overcome all this and joy will return with a beautiful spring and summer! #Cagliari by @*** RT @***\n\nYes, joy will return! Thank you, both, and have a great day! https://*****	Positive
	A month has passed,since Jan.23 Wuhan lockdown,which = two 14-days incubation periods! This could have been used to look for ppl with fever or suspicious symptoms,test them,isolat them,trace their close contacts etc. However,most countries havent done so and wasted a whole month! https://*****	Neutral
	New 1,000-bed facility gets built in record time and is dedicated to treating people infected with coronavirus in Chinese c @*** https://***** https://*****	Negative
8. Cruise	Coronavirus: U.S. to evacuate Americans from cruise ship docked in Japan https://***** via @***: Great news! They deserve to receive best medical care and best accommodations! Kudos to Trump administration!	Positive
	Hey @*** News - Coronavirus: Are cruise ships really 'floating Petri dishes'? https://***** - Stewart Chiron, ""Cruise expert"" sells cruises - hardly an unbiased opinion!	Neutral

	Article: 'A contaminated prison': Scared, angry passengers are trapped on three cruise ships amid coronavirus outbreak - 'A contaminated prison': Scared, angry passengers are trapped on three cruise ships amid coronavirus outbreak\n\nhttps://*****	Negative
9. Hygiene	The best way to protect yourself against #Coronavirus is washing your hand by hand washing regularly, hand sanitizer is also very good but is it not as good as hand washing. On the other hand, eating food which is rich in vitamins is also very important to build up resistant https://*****	Positive
	With both Influenza and Covid19 being a concern at my hospital, this is actually really helpful advice.\n\nIt always feels impersonal and awkward having to forgo a handshake with a patient. https://*****	Neutral
	All this coronavirus scare mongering on @*** is doing my head in. Ban crowds, ban shaking hands, ban this, ban that. \n\nWe should do this every time there's a flu outbreak as well then shouldn't we? \n\n#Getoffthebandwagon #Carryonregardless	Negative
10. People	Coronavirus informative CNN in Ohio. Be proactive and take care of your loved ones, take care of yourselves we are responsible for the wellness of our species. God Bless everyone we can do it!!. It was mentioned money could also be contaminated. Be strong! Good Luck!!! :)	Positive
	Been less scared of the coronavirus ever since it hit the uk lol make it make sense	Neutral
	Before we panic over #Covid_19, here's how Indians die EVERY HOUR:\n70,776 die of non-communicable disease (2017)\n29 die of diabetes (2017)\n24 committed suicide (2017)\n23 die of infectious disease (2017)\n17 die in road accidents (2018)\n0.08 die in terrorism-related acts (2017)	Negative

Table 15 *Sample Tweet from Each Topic and Sentiment for Group 2*

Topic	**Tweet**	**Sentiment**
1. Trump	Pass around some love and kindness - The world and #humanity need it urgently!\nWhat wisdom can you find that is greater than kindness?\nKindness is a vehicle of love and tolerance.\n\n#kindness #love #Respect #patience #Empathy #Integrity \n#Australia #coronavirus #Trump https://*****	Positive
	I totally agree...back in December he was laughing at us all and saying it was the democratic hoax \nSHAME on him\nHow I miss Obama https://*****	Neutral
	My condolences. Donald John Trump should be charged with mass murder. Trump knew of the virus and he refused to tell the people of the USA. Murder Trump. Murder Trump. Lying murder. Lying murder. My he burn in Hell. https://*****	Negative
2. Lockdown	Good time to realign. Good opportunity to look at who I am. Good excuse to practice choosing joy. Time to practice acceptance. Time and freedom to look inward and	Positive

	regroup, albeit somewhat painfully. Time to get good at letting go. Celebrate this Easter with a heart filled w/peace https://***** https://*****	
	This is really good and crucial to the upcoming contested narrative of how ""serious"" the illness is. https://*****	Neutral
	This government through @*** will destroy this f***ing idiotic country and every moron within it who voted for this twat will suffer. F*** you #england f*** you #COVID19 and f*** you @*** you spineless cunt	Negative
3. Death	Happy Free Covid-19 Birthday to me Haha, Thank you lord b'se I wake up to free Oxygen, i really appreciate this other year added to me!!!! with more blessing, Favour, success....Happiest Birthday to Me..... https://*****	Positive
	For all of the government corruption and for a collapsing economy, Lebanon has been able to handle COVID-19 really well. Two months since its first case, Lebanon has reported just 682 coronavirus infections and 22 deaths. Compare that to Cook County, IL https:// ***** https:// *****	Neutral
	#Coronavirus: Global 4,088,393 cases, 281,893 dead; #US 1,327,720 cases, 79,495 dead #UK 31,930 dead #Italy 30,560 dead #Spain 26,621 dead #France 26,383 dead; #Brazil 10757 dead #Belgium 8656 dead #Germany 7560 dead #Iran 6640 dead #Nederlands 5459 dead https:// *****	Negative
4. Update	This is awesome. I love these updates and I bet the rest of the staff/team do as well. What a great way to be there and provide information to your team. It is a scary time and awesome to see the care of one's own so they can best care for others #Paramedics #Leadership #TeamWork https://*****	Positive
	The Pennsylvania Department of Health has announced 1,676 new cases of coronavirus across the state and 78 new deaths. The total for positive cases is 21,655. There have also been 98,498 negative tests to date. #stayconnectedtogether\nhttps://*****	Neutral
	#COVID19: Countries with the highest numbers of infected prisoners:\n\nUSA: 14,502 infected; 172 dead\n\nSyria: 816 infected, 204 dead\n\nChina: 806(+) infected\n\nUK: 311 infected; 16 dead\n\nCanada: 244 infected, 1 dead\n\nMorocco: 210 infected\n\nFrance: 101 infected, 2 dead	Negative
5. Pandemic	You think COVID19 is a pandemic? NO!! Real pandemic is Obesity!!!!!!!!!!	Positive
	83 years old. And homeless. And no place to go. During #COVID19. If you are not outraged, well, I am not quite sure what to say. https://*****	Neutral
	I doubt if it really helps to frame the humanitarian, scientific, and governmental response to the #coronavirus pandemic in terms of wars. Look around. There are so many wars already, everywhere. War against climate. War on terror. War on poverty. War on drugs. War war war.	Negative
6. People	To-day International Laughter day is celebrated all over the world.People laugh & laugh,enjoy & exchange good wishes.Laughing increases immunity & level of positivity.\n Patanjali Yog Samiti celebrated this day performed variant mode of laughter & dedicated to COVID-19 WARRIORS.	Positive

	Germany is so successful against Covid-19 becoz it catches people with no or few symptom very early due to mass testing, hence very less fatalities\n\n#ModiLeadingTheWorld	Neutral
	In a revelation why worrier about coronavirus. Are you scared to die?\nHow many die of car accident, pneumonia,gun killing, knifes killing, flu,cancer,abortion,poisoned,murder by accident, will any body answer my question?	Negative
7. Help	Yesterdays data.\n\nWords describing it would be,\nNOT:\nGood, pleasing, happy, pleasant, optimistic, hopeful.\n\nIt's real and factual data, we hope the wonders of brilliant minds make great achievements.\n\nShare, Care, Pray and Retweet. Keep faith my friends.\n#covid19 #SARSCoV2 https://*****	Positive
	COVID-19 : Community Policing is need of the hour in fight against Corona Pandemic.\n1. Complete lockdown and extensive checking through contact tracing to control it completely like South Korea or China etc. 1/7	Neutral
	Stop the Lies or more will die! The GOP failure is apparent and every american will be affected by their incompetence snd Criminal Negligence! @*** you are pure evil! Your shame will hurt you and all your family. Your tweets will follow you forever. https://*****	Negative
8. Life	I love you but God loves you better. I am your friend but God is your best friend. I can hold your hands but God can hug you. \n\n#AAwords #God #coronavirus #love #bestfriend #hug #positivevibes https://*****	Positive
	Really appreciate all the banks and insurance companies reaching out to let me know they are working hard to make sure I can still pay my bills during this time of crisis. #coronavirus	Neutral
	@*** Please say to China to: FORBID KILLING DOGS, CATS, IMMEDIATELY! \n\nFORBID KILLING ANIMALS\nForbid eating animals\nDo not eat animals: Coronavirus \nDo not be cruel to aninals\ndo not kill animals\nGod sees everything\n\nPOLITICIANS:\nFORBID CRUELTY TO ANIMALS and FORBID KILLING ANIMALS ! https://*****	Negative
9. Hospital	Proud of all health and social care staff providing excellent care to people +caring for each other at same time. Thank you for your courage, care and compassion. We are so lucky to have such amazing staff @*** @*** #proudtobeanurse @*** https://*****	Positive
	'We're Completely Overwhelmed' - Mexico City Hospitals Turn Away Patients As Serious COVID-19 Cases Surge | Zero Hedge https://*****	Neutral
	INTENTIONAL DELAYS by #Mangowanker will cause COVID catastrophe to kill more Americans than World War 2\n\nThe plan:\nDelay, Misdirect, Insult\nMurder blue states cities\nMurder the poor\nCREATE PANIC CHAOS\nDeclare Martial Law\n\n#TrumpMassMurderer\n#TrumpVirus\n\nhttps://*****	Negative
10. News	Heavenly!! Thank you for sharing as art is a spectacular gift from God that we can utilize to help lift spirits and comfort soles along side all of the Great Heroes of our	Positive

	political leaders giving our Great Heroes in our Medical Industry, our Food and Grocery Industry and those https://*****	
	Food delivery services shd stop delivering food unless they are absolutely certain that their riders are not carriers of the #COVID19 virus. @*** @*** @***#COVIDmalaysia	Neutral
	DNC MSM ""its racist to notice horrific violence assaults rapes killings abductions by black men aged 16-32 against blacks whites asians hispanics women children elders disabled. Scream ""black men are victims"" Ignore black violence & crimes. Or else you are racist. https://*****	Negative

To find relationships between sentiment characteristics and discussion topics based on sentiment analysis, the sentiment analysis technique was applied to the negative, neutral, and positive words identified by LDA for each topic for both groups. Among the most frequently used words of selected topics, there are ten negative topics, three positive topics, and 87 neutral topics for group 1, while there are eight negative topics, 14 positive topics, and 78 neutral topics for group 2, as shown in Tables 16 and 17. Therefore, group 2 had more positive top words of selected topics than group 1, while group 1 had more negative topics than group 2 in the top words of selected topics. Table 16 shows the top 10 words of the LDA model for group 1 and Table 17 for group 2.

Table 16 *Top Words of Selected Topics for Group 1*

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
Health	**Wuhan**	**News**	**Market**	**China**
Health	Wuhan	News	Market	China
Trump	China	Update	Stock	*Death***
China	Chinese	Confirmed	Spread	Report
*Emergency***	Doctor	Patient	Vaccine	Number*
Public	Hospital	Live	Global	Toll
Sars	Outbreak	Uk	*Fear***	Country
*Flu***	Video	School	Economy	Iran
Get*	City	Health	Disease	Confirmed
Official	Animal	Rate	Cure	State
Human	People	Tested	Test	Rise
Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
Flight	**Countries**	**Cruise**	**Hygiene**	**People**
China	China	Cruise	Hand*	People
Flight	Day	Ship	Latest	Trump
Impact	Italy	Japan	Daily	*Flu***
Hit	South	News	*Fear***	Going
County	Korea	Passenger	Event	Thing
House	Wuhan	Quarantined	Concern	Time
Business	Quarantine	Test	*Cancel***	*Die***
Airline	Spread	Quarantine	Mask	Pneumonia
Amid	Chinese	American	Amid	*Kill***
Apple	Hospital	People	Stay	*Sick***

Note: Positive and negative words from the LDA model are represented in green (underlined) with one star '' and red (italic) with two stars '**' respectively.*

Table 17 *Top Words of Selected Topics for Group 2*

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
Trump	**Lockdown**	**Death**	**Update**	**Pandemic**
Trump	Lockdown	*Death***	Update	Pandemic
Stay	Country	Time	Thing	Health
Business	Good*	Read	Live	Public
American	Testing	Uk	*Death***	Amid
President	*Crisis***	Number*	Total	Global
Call	Year	Free*	Report	*Risk***
Safe*	India	Rate	Data	Long
Hope*	School	Mask	Confirmed	Change
Vaccine	Hand*	Better*	April	Start
Post	*Flu***	Care	Dr	System
Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
People	**Help**	**Life**	**Hospital**	**News**
People	Help*	China	State	News
Day	Spread	Work	Test	Government
Week	Support*	Life	Patient	Response
Today	Family	Social	Hospital	*Fight***
Daily	Money	Job	Going	Face
Latest	Best*	Medium	Doctor	House
*Died***	Share*	God*	Staff	Medical
*Die***	Human	Man	Child	Leader
Real	Pm	Distancing	Person	Woman
Travel	Friend*	Article	Issue	White

Note: Positive and negative words from the LDA model are represented in green (underlined) with one star '' and red (italic) with two stars '**' respectively.*

Findings Related to The Differences Between Sentiment Characteristics and Discussion Topics:

4. What are the differences between the sentiment characteristics and discussion topics in Covid-19 tweets during the 60 days before and the 60 days after the declaration of the Covid-19 pandemic?

There are significant differences between the two groups. After declaring Covid-19 pandemic rapidly increase in different aspect, such as number of users and places. The group 1 accounted for only 16% of the total dataset while group 2 accounted for 84%. Table 21 shows each the weights of variables in both groups.

Table 18 *Differences Between Two Groups in Description of Covid-19 Tweets*

No. of	G 1	Percentage	G 2	Percentage	Total	Percentage Change
1. Tweets	3,406,055	16%	18,249,229	84%	21,655,283	-68%
2. Unique users	733,219	21%	2,714,759	79%	3,447,978	-58%
3. Verified users	14,745	33%	29,571	67%	44,316	-34%
4. Non-verified users	718,474	21%	2,685,188	79%	3,403,662	-58%
5. Unique places	9,929	25%	29,260	75%	39,189	-50%
6. Unique hashtags	184,763	14%	1,125,396	86%	1,310,159	-72%
7. Retweets	14,693,952	20%	57,514,036	80%	72,207,988	-60%
8. Favorites	40,856,201	18%	186,741,409	82%	227,597,610	-64%

Note: G1= Group 1, G2= Group 2

The Sentiment Analysis showed slightly similar results for the two groups in that the differences are less than 2% for each sentiment characteristic. Group 2 has more positive and neutral tweets than group 1, and group 1 has more negative tweets than group 2. Table 22 shows the sentiment characteristics of the two groups and the differences between.

Table 19 *Sentiment characteristics differences between the two groups*

Sentiment Characteristics	**No. Tweets G1**	**Percentage**	**No. Tweets G2**	**Percentage**	**Percentage change**
Positive Tweets	928,961	27.3%	6,982,729	28.7%	-1.40%
Neutral Tweets	1,114,181	32.7%	5,244,152	33%	-0.30%
Negative Tweets	1,362,913	40%	6,022,348	38.3%	1.70%
Total	**3,406,055**	**100%**	**18,249,229**	**100%**	0%

Note: G1= Group 1, G2= Group 2

The emotional analysis shows that half of the emotions in group 2, anger, anticipation, disgust, sadness, and surprise, had 1% fewer tweets than group 1, in while fear had 2.58% more and negative 6.06% tweets in group 1. However, group 2 had 0.96% more tweets with joy by, 2.56% more with trust, and 6.09% more that were positive. Table 23 shows the differences in emotions between two groups.

Table 20 *Emotional Analysis Differences Between Two Groups*

Emotions	No. Tweets G1	Percentage	No. Tweets G2	Percentage	Percentage Change
Anger	161,461	4.74%	839,986	4.60%	0.14%
Anticipation	391,198	11.49%	2,093,038	11.47%	0.02%
Disgust	134,939	3.96%	699,148	3.83%	0.13%
Fear	349,481	10.26%	1,401,949	7.68%	2.58%
Joy	114,673	3.37%	791,015	4.33%	-0.96%
Sadness	221,059	6.49%	1,067,425	5.85%	0.64%
Surprise	187,331	5.50%	992,612	5.44%	0.06%
Trust	366,203	10.75%	2,429,731	13.31%	-2.56%
Negative	882,087	25.90%	3,620,263	19.84%	6.06%
Positive	597,623	17.55%	4,314,062	23.64%	-6.09%
Total	**3,406,055**	**100.00%**	**18,249,229**	**100.00%**	**0.02%**

Note: G1= Group 1, G2= Group 2

Discussion topics clearly differed between the two groups as shown in Table 24. Regarding the number of tweets for each topic, the distribution of tweets for group 1 was between 6.31% and 18.88%, while for group 2 it was between 6.08% and 31.30%. The percentage change indicates the differences between two groups in topics, such as topic 1 in group 2 was higher than group 1 by 20.81% and topic 10 in group 1 was higher than group 2 by 12.80%. Figure 12 shows the distribution of sentiment characteristics for both groups, illustrating the differences between the two groups in terms of number of tweets.

Table 21 *Differences of No. Tweets on Different LDA Topics For Both Groups*

Topics G1	Total	Percentage of tweets	Topics G2	Total	Percentage of tweets	Percentage Change
1. Health	357,129	10.49%	**1. Trump**	5,711,865	31.30%	**-20.81%**
2. Wuhan	310,091	9.10%	**2. Lockdown**	1,752,644	9.60%	**-0.50%**
3. News	319,097	9.37%	**3. Death**	1,591,210	8.72%	**0.65%**
4. Market	381,493	11.20%	**4. Update**	1,318,173	7.22%	**3.98%**
5. China	311,119	9.13%	**5. Pandemic**	1,780,675	9.76%	**-0.63%**
6. Flight	242,981	7.13%	**6. People**	1,215,728	6.66%	**0.47%**
7. Countries	239,080	7.02%	**7. Help**	1,211,307	6.64%	**0.38%**
8. Cruise	214,944	6.31%	**8. Life**	1,214,364	6.65%	**-0.34%**
9. Hygiene	386,888	11.36%	**9. Hospital**	1,343,759	7.36%	**4.00%**
10. People	643,233	18.88%	**10. News**	1,109,504	6.08%	**12.80%**
Total	**3,406,055**	**100.00%**	**Total**	**18,249,229**	**100.00%**	**100.00%**

Note: G1= Group 1, G2= Group 2

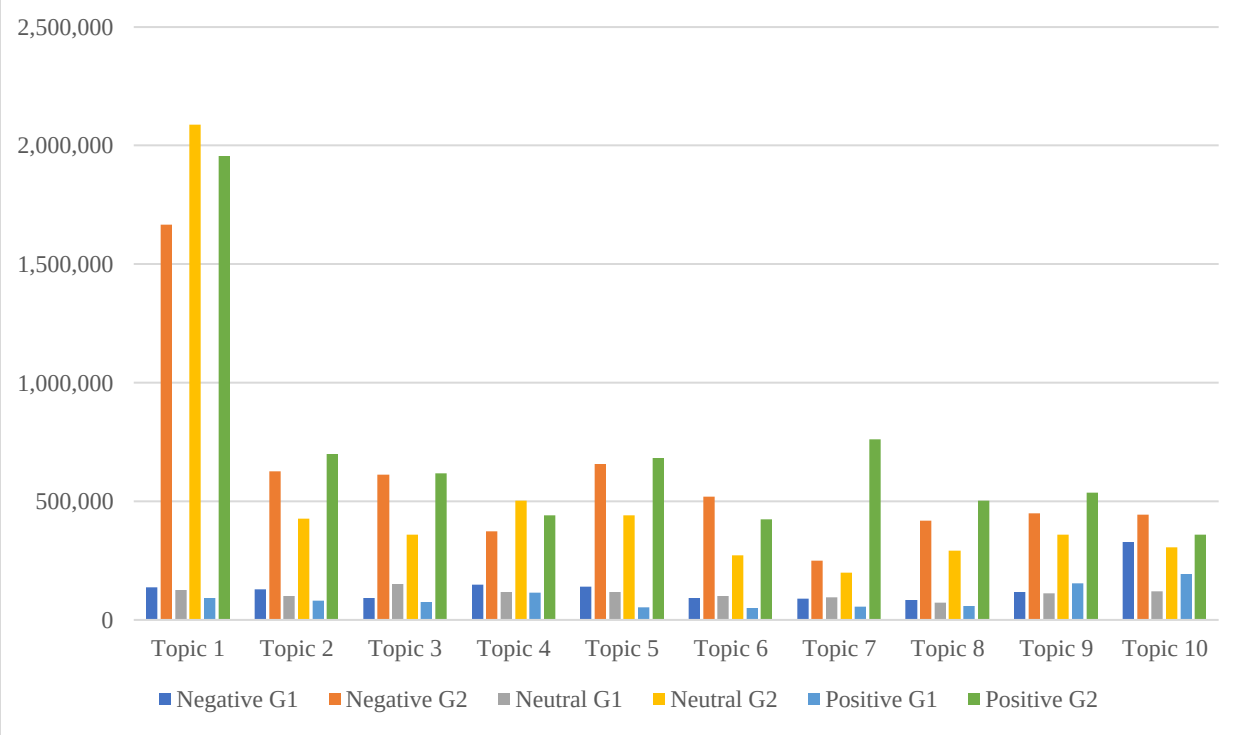

Figure 12. The Distribution of Sentiment Characteristics in Discussion Topics for Both Groups

Inferential Analysis:

The results of Kolmogorov-Smirnov test showed a p-value of 0.00409 with a K-S test statistic (D) of 0.22344 for the sentiment characteristics in discussion topics, which provided good evidence that of non-normal distribution, which means there were differences between the results for sentiment characteristics and for discussion topic. A series of non-parametric tests (e.g. Mann-Whitney U test and Kruskal-Wallis H test) was used to test the differences between the two groups in sentiment characteristics and discussion topics.

The Mann-Whitney U test, which is performed to test for differences between two independent groups (Pallant, 2013), was conducted to test the differences between sentiment characteristics (negative, neutral, and positive) and discussion topics in both groups. The Mann-Whitney U-value indicated a statistically significant difference between the two groups ($\mathrm{U} = 5, \mathrm{Z} = -6.57168, \mathrm{p} = 0.00001 < 0.05$). Moreover, the result of Kruskal-Wallis H test also resulted in a statistically significant difference between the two groups in terms of the expressed sentiment ($\chi^2(1, \mathrm{N} = 60) = 43.2842$) with the p-value 0.00001.

It was revealed that there are significant differences in sentiment characteristics and emotional analysis between tweets posted 60 days before and 60 days after the declaration of the Covid-19 pandemic. There were more negative tweets before than after the declaration, and more positive tweets after the declaration. The results of the emotional analysis showed there were more “fear” and “negative” tweets before than after the Covid-19 pandemic declaration, and more “trust” and “positive” tweets after than before the declaration. Inferential analyses were then conducted to find the differences between the results of sentiment characteristics and discussion topics.

A non-parametric Kolmogorov-Smirnov test showed that the sentiment characteristics and discussion topics during the 60 days before declaration of the Covid-19 pandemic did not differ significantly from normal distribution, but those posted 60 days after the declaration were not normally distributed. The Mann-Whitney U test, and the Kruskal-Wallis H test were also conducted to investigate whether there were any differences between sentiment characteristics and discussion topics in the two groups. The results showed there were significant differences between sentiment characteristics and discussion topics 60 days before and after declaration of the Covid-19 pandemic Table 25 summarizes the inferential analysis results.

Table 22 *Inferential Analysis Summary*

Inferential Analysis	**Test result**	**P-Value**
1. Kolmogorov-Smirnov	K = 0.22344	**P = <0.00409**
2. Mann-Whitney U	U = 5, Z= -6.57168	**P =<0. 00001**
3. Kruskal-Wallis H	H = 43.2842	**P = <0.00001**

Note: Difference is indicated with boldface, and no difference with star ''*

Summary:

Research question one was addressed by the sentiment and emotional analysis techniques to reveal users' characters. Research question two was answered by applying *Latent Dirichlet Allocation* (LDA) to the Covid-19 tweets. The research questions three and four were addressed by inferential analysis to discover the relationships and differences between users' characteristics 60 days before and after the Covid-19 pandemic declaration.

Chapter 5. Discussion:

In this section the major findings reported in Chapter 4 are discussed in reference to each of the research questions individually, followed by a general interpretation of the overall results. The primary objective of this study was to compare the characteristics of tweets posted 60 days before and 60 days after the declaration of the Covid-19 pandemic from the perspective of Social Cognitive Theory (SCT). The theoretical and practical implications of the findings of this study are then explained and future research possibilities are proposed. Following are the four research questions:

1. ***What are the key sentiment characteristics of tweets related to Covid-19 posted during the 60 days before and the 60 days after declaration of the Covid-19 pandemic?***
2. ***What are the key discussion topics emerged from tweets related to Covid-19 during the 60 days before and the 60 days after the declaration the Covid-19 pandemic?***
3. ***What are the relationships between sentiment characteristics and discussion topics in Covid-19 related tweets during the 60 days before and the 60 days after the declaration of the Covid-19 pandemic?***
4. ***What are the differences between the sentiment characteristics and discussion topics in Covid-19 tweets during the 60 days before and the 60 days after the declaration of the Covid-19 pandemic?***

Discussion of Findings Related to The Sentiment Characteristics:

The aim of this research question was to explore the sentiment characteristics and emotional analysis of tweeting related to Covid-19 on Twitter during the 60 days before and the 60 days after the declaration of the Covid-19 pandemic. The findings from analysis of sentiment characteristics and emotional analysis revealed personal factors and behaviors among Twitter

users during the Covid-19 outbreak. Understanding such personal factors is an essential aspect in managing the Covid-19 outbreak or future health crises.

In both groups, over 38% of the tweeting about Covid-19 had negative sentiment characteristics, which was expected. Almost 33% of the tweets about Covid-19 in were neutral, while 27.3% of the tweets in group 1 and 28.7% in group 2 were positive. As noted, the words "negative" and "positive" were removed from the texts to increase the accuracy of the results. As the epidemic progressed, the sentiment characteristics tended to be more positive more information was being reported. The official declaration of a pandemic changed users' sentiment characteristics in how they reacted to the news about the outbreak, such as the number of cases, which is related to knowing as a behavior factor in SCT.

Although the sentiment analysis determined that overall the users produced more negative than neutral or positive tweets, the emotional analysis revealed more categories of users' characteristics, which provide deeper insight into users' personal factors based on SCT. The emotional analysis showed that while more negative emotions concerning Covid-19 were tweeted before the declaration of the pandemic, more positive emotions concerning Covid-19 were tweeted after. Before the declaration, joy was the lowest emotion expressed in only 3.37% of the tweets related to Covid-19, while after the declaration, disgust was the lowest emotional expressed in only 3.83% of the tweets. Table 9 in Chapter 4 presents the categories that were between 3.37% and 25.90% for both groups, showing how the users' emotions changed after the declaration of Covid-19 pandemic and exactly how they reacted to the outbreak. The sentiment and emotional analyses helped clarify the personal factors in SCT, which provide rich information for authorities to consider in health crisis management.

Previous studies identified Java et al. (2007) reported that individuals use Twitter for a variety of reasons such as daily chatter, conversations, sharing information, and reporting news. Which can entail different levels of emotion. Haddi et al. (2013) suggested that pre-processing a text can increase the quality and reliability of sentiment analysis results. Roberts et al. (2012) recommended analyzing texts with different types of emotions to gain a broader sense of the feelings among users. Various keywords related to public awareness were applied in the emotional analysis. The results concerning public emotion showed that there were more negative than positive emotions in group 1 and more positive than negative in group 2, with "trust" as the most negative word in both groups (Dubey & Tripathi, 2020; Kleinberg et al., 2020). In previous pandemics, negative sentiments were generally prevalent in social media (Mamidi et al., 2019; Zhang et al., 2019).

It has been found in some studies employing sentiment and emotional analysis of Twitter that most tweets about Covid-19 were negative (Abd-Alrazaq et al., 2020; Chakraborty et al., 2020; Gupta et al., 2020). It was found that sentiment categories were linked to different issues concerning Covid-19 crisis management, such as fears related to shortages of COVID-19 tests and medical supplies and anger ranging from xenophobia to resentment of stay-at-home mandates, and sadness related to losing friends and family members. Expressions of joy included words of gratitude and good health (Lwin et al., 2020; Sesagiri Raamkumar et al., 2020). As in this study, Chakraborty et al. (2020) found that the positive and neutral tweets were more frequent than negative tweets; however, retweets were mostly negative. However, in the previous literature, in most studies not as many tweets for as long a period of time as in this study were collected, so they failed to provide as broad a picture of the sentiment characteristics of users, including both personal and public factors of the health crisis. Also, previous studies of

social media activity related to the Covid-19 pandemic were not based on Social Cognitive Theory.

In this study, both analysis of sentiment characteristics and emotional analysis were used to investigate Twitter users' personal factors based on SCT. The significant result of these analyses is that emotions regarding the Covid-19 pandemic were found to be more negative in the 60 days before the declaration of the pandemic and positive tweets in the 60 days after. This finding was consistent with that of Gupta's (2020) sentiment and emotional analysis showing negative tweets about the Covid-19 pandemic.

Discussion of Findings Related to The Discussion Topics:

The discussion topics demonstrate Twitter users' focus when discussing and reacting to the Covid-19 pandemic 60 days before 60 days after the pandemic was declared. The that emerged revealed the environmental factors among users. This kind of information can help public health agencies and authorities understand public concerns as well as what public health information is resonating in among users in Twitter. To reduce the spread of Covid-19, Public health agencies and governments in North America have focused their messaging on encouraging hand hygiene, social distancing, and staying home. This messaging was echoed by both groups in this study in discussions about hand washing, staying home, mask-wearing and social distancing in both groups. According to SCT as employed in this study, such messaging on environmental factors can have a positive or negative impact on Twitter users

The LDA model was employed for topic modeling using machine learning to discover what the Twitter users discussed the 60 days before and the 60 days after the pandemic was declared. The findings of the topic modeling reveal that the users discussed various topics in Twitter that related to Covid-19 pandemic. The discussion topics and terms were represented in

as hashtags, which the LDA model showed (Table 5 in Chapter 4). Before the pandemic was declared, users discussed how and where Covid-19 started to spread as well as its incidence among countries, effect on economies, Covid-19 symptoms, and numbers of cases and deaths. Among the most discussed subjects among Twitter users in the 60 days before the declaration of the pandemic were China, people, health, Wuhan, news, Trump, Chinese, confirmed, spread, hospital, fear, test, quarantine, and flu. When governments first declared Covid-19 as a pandemic, the topics were different, such as the origin of Covid-19 and which countries had high numbers of cases. Also, public health agencies messaged people on ways to avoid sickness. Different aspects of the economy were also discussed by users in Twitter in the 60 days before declaration of the Covid-19 pandemic. Table 14 in Chapter 4 shows examples of tweets on various topics discussed by group 1 before the declaration and Table 15 shows examples for group 2. Users started to discuss various topics that related to their current situation when information became available from different sources, such as numbers of cases and deaths and governments restrictions. Nevertheless, other news and China topics were continuous discussion topics in group 1. Although several terms were repeatedly used in group 1, only one term was repeatedly used in group 2, which was death. The discussed topics included updates of the number of cases and deaths, governments' actions, and health organizations. Users focused on deaths under two topics. The discussion topics and terms changed from the spread of Covid-19 to governmental restrictions and news.

Similar patterns of topics are found in studies of Facebook and Twitter posts using natural language processing (NLP) and deep learning techniques. Declaring Covid-19 as a pandemic has stimulated public concern globally. Among the 10 main topics across both groups were (1) reports of numbers of cases and deaths, (2) new governmental restrictions, (3) news

reports, and (4) health guidance. Major themes during previous disease outbreaks that were similar to those found in this study were reported by several researchers (Abd-Alrazaq et al., 2020; Dubey & Tripathi, 2020; Gupta et al., 2020; Nussbaumer-Streit et al., 2020; Sesagiri Raamkumar et al., 2020; Shah et al., 2020). These themes included prevention and control procedures, including quarantine, as well as reports on confirmed cases and medical treatments.

The findings of this study show that the discussion topics concerning the COVID-19 pandemic in both groups were multifaceted, and that the public was actively seeking and sharing information about the situation on the social media. Also, in this study some changes between two groups were found in the patterns of discussion topics, indicating how the official declaration of Covid-19 as a pandemic changed the discussion topic in various ways, such as increased interest in news reports and governmental restrictions, which is known as the environmental effect on users' behavior in SCT that. It is important for public health agencies and authorities to take serious action to understand the public's major interests and concerns to inform decision-making and management of future health crises.

Discussion of Findings Related to The Relationship Between Sentiment Characteristics and Discussion Topics:

As discussed, previous researchers have investigated sentiment characteristics and discussion topics on Twitter in various fields. However, a few studies have addressed the relationship between sentiment characteristics and discussion topics by applying different analysis techniques, such as joint sentiment topics (JST). In this study two datasets representing the 60 days before and the 60 days after the declaration of the Covid-19 pandemic were compared, and the findings related to sentiment characteristics and discussion topics were shown on matrix tables for group 1 (Table 12) and for group 2 (Table 13). Then sentiment analysis was

applied to the topic modeling results to discover the sentiment characteristics within discussion topics.

Topics' sentiment analysis the VADER model in Python was employed to analyze sentiments within topics. Two tables showed the combined results including the most essential findings for both groups. For group 1, as shown in Table 12, the largest numbers of tweets under topic 10 were negative and positive, which included tweets about people, sickness, and death. For group 2, it can be seen in Table 13 that the largest numbers of tweets under topic 1 were neutral, positive, and negative, which included tweets about Trump, business, and America.

Interestingly, this matrix can be linked that with the results of sentiment topics. In group 1, as can be seen in Table 16 showing top words of selected topics, for group 1, topic 10 (People) has four negative terms in the finding of sentiment topics, which is the most frequently mentioned negative topic in table 12. Topic 3 (News) for group 1 has the second largest number of neutral tweets, and all the terms in Table 16 are neutral. For group 2, while topic 1 (Trump) has the largest number of neutral tweets in Table 12, it has only two positive terms with eight neutral terms in the findings of sentiment topics. Also, as strong evidence of the sentiment topic linkage in the matrix, topic 7 (Help) is the most frequently mentioned positive topic in Table 13, and it has five positive terms in Table 17. The result of sentiment topics as shown in the matrix table indicates a significant relationship between sentiment and topics when sentiment topics analysis was performed on the terms.

One example of a tweet showing a relationship between sentiment and topic 3 (News) in group 1 was a comment a about cryptocurrency with positive characteristics: "*Dear Crypto Friends, \n\nOne of our Top authors has been affected by the coronavirus, Let's hope and pray for her recovery, sharing with you her #BTC Adress and i'm sure she will appreciate any of your*

*#donations \nbtc: *******f\n\nThank you,\nThe TCAT team.*" Also, there was a negative tweet on topic 4 (market): "*Oil drops the most since July on fears coronavirus will hit growth https://****** https://*******". These tweets were about news reports and the market, for which the relationships results show a strong significant relationship between sentiment and topics.

In group 2, the strongest relationships between topic 6 (people) and topic 1 (Trump) were with neutral tweets, such as the following: "*I got a little problem with Trump on that. No clinical trials on that malaria drug being off-label prescribed for Covid 19. It's being pushed by Pres Trump & now many others. Some bad side-effects. One must be careful. FDA approved an off-label malaria drug to treat COVID 19. https://******j,*" This tweet has a relationship with another tweet in neutral topic 6: "*This is the tremendous success .@****** and .@***** are pushing. They're always pushing something. One week it's drugs and the next week it's *****. #MAGA https://******. "*

Discussion of Findings Related to The Differences Between Sentiment Characteristics and Discussion Topics:

Differences between the sentiment characteristics and discussion topics were evident when no differences were between a few topics and sentiment characteristics. To address this research question, the different percentages between the two groups were calculated as shown in Table 23. The average percentage was found to be 58%, which means that group 2 had a higher number of interactions than group 1.

After the declaration of the Covid-19 pandemic, group 2 had different reasons for increased information than group 1 even though both represented the same range of time. However, governments took different actions to contain the spread of Covid-19 after the WHO

declared the Covid-19 pandemic, such as closing borders and issuing stay-at-home orders. Consequently, some factors of the Covid-19 crisis, such as news reports, government actions, and numbers of cases and deaths (environmental factors in SCT) made users become active and interact more in Twitter than before the declaration to share information about the Covid-19 crisis.

In the sentiment and emotional analysis, the percentage of negative tweets in group 1 was larger than in group 2, while the percentage of positive tweets in group 2 was larger than in group 1. The emotional analysis also showed a slight difference between the groups, except for fear, trust, negative, and positive tweets. The percentages of fear and negative tweets in group 1 were larger than in group 2, while the percentage of positive tweets in group 1was larger than in group 2. Environmental factors affected personal factors in tweets expressing fear and trust and in positive and negative tweets. Moreover, a few discussion topics and terms were similar in both groups, but they had many different topics and terms. Interestingly, the top words of selected topics in group 2 had more positive terms than group 1, while group 1 had more negative terms than group 2.

The statistical analyses revealed differences between sentiment characteristics and discussion topics before and the declaration Covid-19 pandemic. Table 27 summarized all the results of tests, which appear that there are differences between sentiment characteristics and discussion topics in both groups. Results of the sentiment characteristics demonstrated that overall, attitudes conveyed by users tweets in Twitter were mildly negative in group 1 and positive in group 2.

Implications:

Theoretical implications:

The theoretical implications of this study lie in the use of Social Cognitive Theory to guide an investigation of users' behaviors on Twitter in early stages of the Covid-19 outbreak. Whereas in previous studies of the characteristics of tweets related to the Covid-19 pandemic small to medium sized datasets covering short periods of time, in this study a large-scale dataset of all Covid-related Twitter chatter in English covering two critical periods of time, just before and just after the official declaration of the pandemic, was used. Based on SCT, sentiment analysis, topic modeling, and other statistical analyses were applied to this comprehensive data to explore the personal and environmental factors affecting people's reactions to the health crisis.

Methodological Implications:

Sentiment analysis and topic modeling have been extensively applied in research on social media, especially Twitter. These methods provided distinctive ways to explore the impact of the Covid-19 on the reactions and behaviors of a broad range of Twitter communities. In this study, machine learning techniques facilitated fast, convenient, and accurate analysis of the large-scale dataset. Sentiment analysis and topic modeling were conducted to seek the patterns of users' behaviors during the early stage of Covid-19 pandemic. Altogether, seven distinct statistical methods to were applied to discover the relationships and differences between sentiment characteristics and discussion topics of information exchange on Twitter before and after the declaration of the Covid-19 pandemic. This methodology can be used to explore other social media platforms focusing on other health crises.

Practical Implications:

The results of this study demonstrate the value of Twitter data to help policymakers should explore and monitor public awareness and emotions regarding the COVID-19 pandemic and other health and safety crises. The levels of public awareness are found to be dynamic as evidenced by the two awareness peaks identified in this study over just a few months, which is a crucial aspect to understand for crisis management. Users shared both information and misinformation via he social media platform during the different stages of the disease, which can help authorities maintain a flow of accurate information. The finding that people tended to express negative emotions 60 days before and positive emotions 60 days after the declaration of the pandemic suggests the importance of providing as much certainty as possible. Users expressed great fear before they were sure of what was happening and trust after the declaring of pandemic had clarified the situation. Overall, negativity and fear characteristics were prevalent before and the positive and trust characteristics increased after the Covid-19 pandemic was declared.

The findings of this study can help policymakers understand how news reports and governmental actions can have positive or negative impacts on communities, which is useful input to help governments clearly communicate important and accurate information regarding COVID-19 and any other public emergency to their constituents. Misinformation or "fake news" on the internet and social media can create mass panic and result in negative actions concerning whatever crisis must be met with unity and determination.

In the case of the Covid-19 pandemic, Government agencies should synchronize and monitor the flow of accurate information and combat "fake news" to diminish fear and create confidence in measures being taken. The findings that the positive emotion and trust increased

after the declaration of the Covid-19 pandemic, while negativity and fear were prevalent before suggest that policymakers should have taken countermeasures to mitigate these emotions before declaring the Covid-19 pandemic by building national surveillance systems to monitor social media platforms to understand the emotions within the community and employ decision-support committees accordingly.

In this study, topics derived from user-generated tweets about the Covid-19 pandemic were investigated. The four themes (health concerns, public health interventions, controlling the pandemic, and the economy) reflected the issues that users were concerned about with regard to the pandemic. These topics can guide the design of a more proactive public health presence on social media by establishing decision-support committees and building national surveillance systems to assess social media platforms. Understanding the relationships and differences among these key topics can help policy makers identify links between topics and efficiently address two or more issues at one time.

Summary:

The results of this study were compared to the findings from prior studies. The findings of this study could increase policy makers' understanding of how to manage and communicate information and deal with communities though social media. The theoretical and practical implications of the study were discussed. Different methods were applied to explore users' social media behaviors during the period of the Covid-19 pandemic, which can provide a foundation for research on the role of social media in other health crises on. The practical implications of the results from this study can support health organizations, governments, and policy makers to make appropriate decisions on health crisis management.

Chapter 6. Conclusion:

This chapter summarizes the research questions and related primary findings from the Results and Discussion chapters. The limitations of this study and future directions are also discussed in this chapter.

Key of Research Findings:

The purpose of this dissertation study was to reveal and compare the characteristics of Twitter communications posted 60 days before and 60 days after the WHO's declaration of the Covid-19 pandemic. Data for this study were obtained from the Panacea Lab and prepared for multiple analyses, which included sentiment analysis, emotional analysis, the LDA model, and inferential analysis. The users' characteristics were more negative, fear, and sadness, less positive, joy, and trust before declaration of the Covid-19 pandemic. After declaration of the Covid-19 pandemic, users' characteristics were more positive, joy, and trust, and less negative, fear, and sadness.

The results suggest the need for and feasibility of establishing a fast and low-human-effort surveillance system for monitoring people's attitudes towards policies and policymakers during a health or other crisis using Twitter or other social media platforms. This study demonstrated that monitoring tweets was an effective way to determine public response to the Covid-19 pandemic and government efforts to control its spread in order to maintain appropriate levels of control, which is applicable not only to the Covid-19 pandemic but also to other crises for policymakers. Decisions made by governments or global organizations may change individuals' attitudes in positive or negative ways. For instance, the tweets analysis in this study showed that many users changed their attitudes towards the stay-at-home order due to the negative effects of this action. The people responsible for decision-making might have benefited

from using social media platforms as a surveillance system to monitor attitudes within communities. The answers to the four research are briefly summarized as follows:

1. ***What are the key sentiment characteristics of tweets related to Covid-19 posted during the 60 days before and the 60 days after declaration of the Covid-19 pandemic?***

The users tended to exhibit negative characteristics during the 60 days before the declaration of the Covid-19 pandemic and positive characteristics during the 60 days after the declaration. The emotional analysis revealed that tweets with anticipation characteristics were most numerous during the 60 days before the declaration of the Covid-19 pandemic, and those with trust characteristics were most numerous after the declaration. Thus, the WHO's official declaration of the Covid-19 pandemic and subsequent governmental action impacted individuals' emotional reactions.

2. ***What are the key discussion topics emerged from tweets related to Covid-19 during the 60 days before and the 60 days after the declaration the Covid-19 pandemic?***

To answer this research question, the topic modeling method, Latent Dirichlet Allocation (LDA), was applied to tweets posted by both group 1 (pre- pandemic declaration) and group 2 (post-pandemic declaration). The discussion topics most frequently posted by group 1were as follows: Health, Wuhan, News, Market, China, Flight, Countries, Cruise, Hygiene, and People. Those posted most frequently by group 2 were as follows: (Trump, Lockdown, Death, Update, Pandemic, People, Help, China, Hospital, and News). This LDA method showed that each group had particular discussion topics, but there were some similarities between them.

3. ***What are the relationships between sentiment characteristics and discussion topics in Covid-19 related tweets during the 60 days before and the 60 days after the declaration of the Covid-19 pandemic?***

There was a significant relationship between sentiment characteristics and discussion topics in Covid-19 related tweets as disclosed by applying different methods. The matrix tables provide a comprehensive idea of the classification of tweets based on topics and sentiments. Applying sentiment analysis to the LDA model provided strong evidence of the relationships between them.

4. ***What are the differences between the sentiment characteristics and discussion topics in Covid-19 tweets during the 60 days before and the 60 days after the declaration of the Covid-19 pandemic?***

The percentages of the differences between the sentiment characteristics and discussion topics and of the differences between the groups 1 and 2 were calculated and inferential analysis was applied (Kolmogorov-Smirnov test). The results showed there were differences between the sentiment characteristics and discussion topics in tweets during the 60 days before and the 60 days after the declaration of the Covid-19 pandemic.

Limitations:

The first limitation this study is related to sampling and data collection. Covid-19 related tweets in English were the only data collected from social media platforms. Many social media platforms are used globally, some of which are banned in some countries. The sampled Covid-19 English tweets may not represent all of Covid-19 social media communications related to the pandemic as there were many different languages in the original dataset. Moreover, due to the massive size of the data, the data collection periods were confined to 60 days before and 60 days after the declaration of the Covid-19 pandemic. These lengths of time might not show a full picture of users' behaviors.

Another limitation is related to the quantitative research methods used to investigate the users' Twitter behaviors in the early stages of the Covid-19 pandemic. The meanings of tweets might have been obscured by the use of slang, which might have affected the result. The "Syuzhet" library used in the emotional analysis has only eight different emotions and does not account for sarcasm and irony in emotional expressions. However, in this study no users were interviewed about their Covid-19 related tweets. Such interviews may be conducted in future qualitative research.

Future Directions:

The methods used in this study of the characteristics of communications on Twitter during the early stage of the Covid-19 pandemic can be applied to different social media platforms such as Facebook and the results compared to those of this study. As many health organizations and authorities use Twitter, future researchers may capture a bigger picture of users' reactions to the declaration of the Covid-19 pandemic and its consequences, such as stay-at-home home orders by adapting the qualitative methodology. Also, in this study "VADER" was used in the sentiment analysis, "Syuzhet"in the emotional analysis, and 'sklearn' in LDA topic modeling. There are other machine learning models that may provide interesting results, such as 'Textblob', lexicon-based algorithms, and "Gensim."

As this study focused on the users' tweeting behaviors in the early stage of the Covid-19 pandemic on Twitter using quantitative methods, qualitative methods might be used in future studies. Future researchers might focus on enhancing our method by classifying the tweets based on places and languages and search different government-issued orders. Network analysis may be used in future analyses to examine the connections among users, which might show if there were any manipulations of tweets. Future researchers might also interview active Twitter users to

understand their characteristics during the early stage of the Covid-19 pandemic and compare the results with our findings.

Reference:

@JacquelineZote. (2019). 65 Social Media Statistics to Bookmark in 2019.

Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., & Shah, Z. (2020). Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study. *Journal of medical Internet research*, *22*(4), e19016-e19016. https://doi.org/10.2196/19016

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., & Liu, B. (2011). Predicting flu trends using twitter data. 2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS),

Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008). Proceedings of the 2008 International Conference on web search and data mining. In (pp. 183-194).

Al-Rakhami, M. S., & Al-Amri, A. M. (2020). Lies Kill, Facts Save: Detecting COVID-19 Misinformation in Twitter. *IEEE access*, *8*, 155961-155970. https://doi.org/10.1109/ACCESS.2020.3019600

Andersen, P. (2007). *What is Web 2.0?: ideas, technologies and implications for education* (Vol. 1). JISC Bristol.

Arafat, S. M. Y., Kar, S. K., Marthoenis, M., Sharma, P., Hoque Apu, E., & Kabir, R. (2020). Psychological underpinning of panic buying during pandemic (COVID-19). *Psychiatry research*, *289*, 113061-113061. https://doi.org/10.1016/j.psychres.2020.113061

Asiri, E., Khalifa, M., Shabir, S.-A., Hossain, M. N., Iqbal, U., & Househ, M. (2017). Sharing sensitive health information through social media in the Arab world. *International Journal for Quality in Health Care*, *29*(1), 68-74. https://doi.org/10.1093/intqhc/mzw137

Babbie, E. R. (2016). *The practice of social research* (Fourteenth edition. ed.). Cengage Learning.

Bajema, K. L., Oster, A. M., McGovern, O. L., Lindstrom, S., Stenger, M. R., Anderson, T. C., . . . Oliver, S. E. (2020). Persons Evaluated for 2019 Novel Coronavirus - United States, January 2020. *MMWR. Morbidity and mortality weekly report*, *69*(6), 166-170. https://doi.org/10.15585/mmwr.mm6906e1

Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., & Chowell, G. (2020). A large-scale COVID-19 Twitter chatter dataset for open scientific research--an international collaboration. *arXiv preprint arXiv:2004.03688*.

Bandura, A. (1986). *Social foundations of thought and action : a social cognitive theory*. Englewood Cliffs, N.J. : Prentice-Hall.

Barsevick, A. M., & Johnson, J. E. (1990). Preference for information and involvement, information seeking and emotional responses of women undergoing colposcopy. *Research in nursing & health*, *13*(1), 1-7.

Basant, A., Namita, M., Pooja, B., & Sonal, G. (2015). Sentiment Analysis Using Common-Sense and Context Information. *Computational intelligence and neuroscience*, *2015*, 715730-715739. https://doi.org/10.1155/2015/715730

Basch, C., MacLean, S., Romero, R.-A., & Ethan, D. (2018). Health Information Seeking Behavior Among College Students. *The Publication for Health Promotion and Disease Prevention*, *43*(6), 1094-1099. https://doi.org/10.1007/s10900-018-0526-9

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Borgatti, S. P. (2013). *Analyzing social networks*. Los Angeles ; London : SAGE Publications.

Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In (pp. 1-10).

Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW). *Gainesville, FL: The NIMH Center for the Study of Emotion and Attention, University of Florida*.

Briones, R., Nan, X., Madden, K., & Waks, L. (2012). When vaccines go viral: an analysis of HPV vaccine coverage on YouTube. *Health communication*, *27*(5), 478-485.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc.

Cain, M. M., Sarasohn-Kahn, J., & Wayne, J. C. (2000). Health e-people: the online consumer experience. *Institute for the Future*, 1-73.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In (pp. 675-684).

Centers for Disease Control and Prevention. (2020). In.

Chae, B. (2015). Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research. *International journal of production economics*, *165*, 247-259. https://doi.org/10.1016/j.ijpe.2014.12.037

Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R., & Hassanien, A. E. (2020). Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. *Applied soft computing*, *97*, 106754-106754. https://doi.org/10.1016/j.asoc.2020.106754

Chu, S.-C. (2011). Viral advertising in social media: Participation in Facebook groups and responses among college-aged users. *Journal of interactive advertising*, *12*(1), 30-43.

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., . . . Scala, A. (2020). The COVID-19 social media infodemic. *Scientific reports*, *10*(1), 16598-16598. https://doi.org/10.1038/s41598-020-73510-5

Conley, V. M. (1998). Beyond Knowledge Deficit to a Proposal for Information-Seeking Behaviors. *International Journal of Nursing Terminologies and Classifications*, *9*, 129-135.

Corbo-Richert, B., Caty, S., & Barnes, C. M. (1993). Coping behaviors of children hospitalized for cardiac surgery: A secondary analysis. *Maternal-child nursing journal*.

Cotten, S. R., & Gupta, S. S. (2004). Characteristics of online and offline health information seekers and factors that discriminate between them. *Social science & medicine*, *59*(9), 1795-1806.

Dietz-Uhler, B., & Bishop-Clark, C. (2001). The use of computer-mediated communication to enhance subsequent face-to-face discussions. *Computers in Human Behavior*, *17*(3), 269-283.

Dubey, A. D., & Tripathi, S. (2020). Analysing the Sentiments towards Work-From-Home Experience during COVID-19 Pandemic. *Journal of Innovation Management*, *8*(1). https://doi.org/10.24840/2183-0606_008.001_0003

E Hilliard, M., M Sparling, K., Hitchcock, J., K Oser, T., & K Hood, K. (2015). The emerging diabetes online community. *Current diabetes reviews*, *11*(4), 261-272.

El Ouirdi, M., Segers, J., El Ouirdi, A., & Pais, I. (2015). Predictors of job seekers' self-disclosure on social media. *Computers in Human Behavior*, *53*, 1-12. https://doi.org/10.1016/j.chb.2015.06.039

Fan, W., & Gordon, M. (2014). The power of social media analytics. *Communications of the ACM*, *57*(6), 74-81. https://doi.org/10.1145/2602574

FDA. (2020). *Pfizer-BioNTech COVID-19 Vaccine*. Retrieved March, 16 from https://www.fda.gov/emergency-preparedness-and-response/coronavirus-disease-2019-covid-19/pfizer-biontech-covid-19-vaccine

Feinerer, I., Hornik, K., & Feinerer, M. I. (2015). Package 'tm'. *Corpus*, *10*(1).

FICCI. (2020). Impact of COVID-19 on Indian Economy: COVID-19 Mitigation Measures Taken by Indian Companies. *FICCI Studies and Surveys*.

Garcia, K., & Berton, L. (2021). Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Applied soft computing*, *101*, 107057-107057. https://doi.org/10.1016/j.asoc.2020.107057

George, N., Britto, D. R., Krishnan, V., Dass, L. M., Prasant, H., & Aravindhan, V. (2018). Assessment of hashtag (#) campaigns aimed at health awareness in social media. *Journal of education and health promotion*, *7*.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012-1014.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, *1*(12), 2009.

Gollop, C. J. (1997). Health information-seeking behavior and older African American women. *Bulletin of the Medical Library Association*, *85*(2), 141.

Gostin, L. O., & Wiley, L. F. (2020). Governmental Public Health Powers During the COVID-19 Pandemic: Stay-at-home Orders, Business Closures, and Travel Restrictions. *JAMA : the journal of the American Medical Association*, *323*(21), 2137-2138. https://doi.org/10.1001/jama.2020.5460

Grajales Iii, F. J., Sheps, S., Ho, K., Novak-Lauscher, H., & Eysenbach, G. (2014). Social Media: A Review and Tutorial of Applications in Medicine and Health Care. *Journal of Medical Internet Research*, *16*(2), []. https://doi.org/10.2196/jmir.2912

Gravetter, F. J. (2012). *Research methods for the behavioral sciences* (Fourth edition. ed.). Australia ; Belmont, CA : Wadsworth.

Gupta, M., Bansal, A., Jain, B., Rochelle, J., Oak, A., & Jalali, M. S. (2021). Whether the weather will help us weather the COVID-19 pandemic: Using machine learning to measure twitter users' perceptions. *International journal of medical informatics (Shannon, Ireland)*, *145*, 104340-104340. https://doi.org/10.1016/j.ijmedinf.2020.104340

Gupta, R. K., Vishwanath, A., & Yang, Y. (2020). Global Reactions to COVID-19 on Twitter: A Labelled Dataset with Latent Topic, Sentiment and Emotion Attributes.

Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, *17*, 26-32.

Hanah, E. (2019). Business so Tweet: how Twitter drives business growth beyond marketing. *The Business Journal - Central New York*, *33*(12), 10-10.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining (adaptive computation and machine learning)*. MIT Press.

Hazel, B. (2020). COVID-19 pushes social issues to ESG forefront; Racial justice calls add to concern for workers during deadly pandemic. *Pensions & investments (1990)*, *48*(16), 1.

Heinonen, K. (2011). Consumer activity in social media: Managerial approaches to consumers' social media behavior. *Journal of Consumer Behaviour*, *10*(6), 356-364. https://doi.org/10.1002/cb.376

HIM Careers - Health Information 101. (2019). https://www.ahima.org/careers/healthinfo

Hughes, E. (2016). Can Twitter improve your health? An analysis of alcohol consumption guidelines on Twitter. *Health Information & Libraries Journal*, *33*(1), 77-81. https://doi.org/10.1111/hir.12133

Hunt, D., Koteyko, N., & Gunter, B. (2015). UK policy on social networking sites and online health: From informed patient to informed consumer? *Digital health*, *1*, 2055207615592513.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media,

Impact of COVID-19 on Indian Economy: COVID-19 Mitigation Measures Taken by Indian Companies. (2020). *FICCI Studies and Surveys*.

Jain, S. (2012). 40 Most Popular Social Networking Sites of the World.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis,

Jianqiang, Z., & Xiaolin, G. (2017). Comparison research on text pre-processing methods on twitter sentiment analysis. *IEEE Access*, *5*, 2870-2879.

Jockers, M. (2017). *"Syuzhet Package in R."* Retrieved 10, March from https://www.rdocumentation.org/packages/syuzhet/versions/1.0.4

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Kang, M., Ahn, J., & Lee, K. (2018). Opinion mining using ensemble text hidden Markov models for text classification. *Expert Systems with Applications*, *94*, 218-227.

Kapp, J. M. P., Hensel, B. P., & Schnoring, K. T. B. S. (2015). Is Twitter a forum for disseminating research to health policy makers? *Annals of epidemiology*, *25*(12), 883-887. https://doi.org/10.1016/j.annepidem.2015.09.002

Kim, A. E., Hansen, H. M., Murphy, J., Richards, A. K., Duke, J., & Allen, J. A. (2013). Methodological considerations in analyzing Twitter data. *Journal of the National Cancer Institute Monographs*, *2013*(47), 140-146.

Kim, K.-S., Sin, S.-C. J., & Yoo-Lee, E. Y. (2014). Undergraduates' Use of Social Media as Information Sources. *College & Research Libraries*, *75*(4), 442. https://doi.org/10.5860/crl.75.4.442

Kleinberg, B., van der Vegt, I., & Mozes, M. (2020). Measuring Emotions in the COVID-19 Real World Worry Dataset.

Knell, G., Robertson, M. C., Dooley, E. E., Burford, K., & Mendez, K. S. (2020). Health Behavior Changes During COVID-19 Pandemic and Subsequent "Stay-at-Home" Orders. *International journal of environmental research and public health*, *17*(17), 6268. https://doi.org/10.3390/ijerph17176268

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of healthcare information management*, *19*(2), 65.

Kostkova, P., de Quincey, E., & Jawaheer, G. (2010). The potential of social networks for early warning nad outbreak detection systems: the swine flu Twitter study. *International Journal of Infectious Diseases*, *14*, e384-e385. https://doi.org/10.1016/j.ijid.2010.02.475

Kratzke, C., Wilson, S., & Vilchis, H. (2013). Reaching Rural Women: Breast Cancer Prevention Information Seeking Behaviors and Interest in Internet, Cell Phone, and Text Use. *The*

Publication for Health Promotion and Disease Prevention, *38*(1), 54-61. https://doi.org/10.1007/s10900-012-9579-3

Kumar, A., & Sebastian, T. M. (2012). Sentiment analysis on twitter. *International Journal of Computer Science Issues (IJCSI)*, *9*(4), 372.

Kumar, S. (2013). *Twitter data analytics*. https://doi.org/10.1007/978-1-4614-9372-3

Lai, L., & Turban, E. (2008). Groups Formation and Operations in the Web 2.0 Environment and Social Networks. *Group Decision and Negotiation*, *17*(5), 387-402. https://doi.org/10.1007/s10726-008-9113-2

Lambert, S. D., & Loiselle, C. G. (2007). Health information—seeking behavior. *Qualitative health research*, *17*(8), 1006-1019.

Lenz, E. R. (1984). Information seeking: a component of client decisions and health behavior. *Advances in Nursing Science*.

Liang, H., Fung, I. C.-H., Tse, Z. T. H., Yin, J., Chan, C.-H., Pechta, L. E., . . . Fu, K.-W. (2019). How did Ebola information spread on twitter: broadcasting or viral spreading? *BMC public health*, *19*(1), 438-438. https://doi.org/10.1186/s12889-019-6747-8

Liddy, E. D. (2000). Text Mining. *Bulletin of the American Society for Information Science and Technology*, *27*(1), 13-14. https://doi.org/10.1002/bult.184

Lin, H.-C., & Chang, C.-M. (2018). What motivates health information exchange in social media? The roles of the social cognitive theory and perceived interactivity. *Information & Management*.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

Lober, W. B., & Flowers, J. L. (2011). Consumer Empowerment in Health Care Amid the Internet and Social Media. *Seminars in Oncology Nursing*, *27*(3), 169-182. https://doi.org/10.1016/j.soncn.2011.04.002

Loiselle, C. G. (1996). Self-evaluation and health information-seeking: a study of self-assessment and self-protection motives.

Lu, K., & Wolfram, D. (2012). Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches. *Journal of the American Society for Information Science and Technology*, *63*(10), 1973-1986. https://doi.org/10.1002/asi.22628

Lwin, M. O., Lu, J., Sheldenkar, A., Schulz, P. J., Shin, W., Gupta, R., & Yang, Y. (2020). Global Sentiments Surrounding the COVID-19 Pandemic on Twitter: Analysis of Twitter Trends. *JMIR public health and surveillance*, *6*(2), e19447-e19447. https://doi.org/10.2196/19447

Lyu, J. C., & Luli, G. K. (2021). Understanding the Public Discussion About the Centers for Disease Control and Prevention During the COVID-19 Pandemic Using Twitter Data: Text Mining Analysis Study. *Journal of medical Internet research*, *23*(2), e25108-e25108. https://doi.org/10.2196/25108

Mabey, B. (2015). *pyLDAvis*. Retrieved 09, 17 from https://pyldavis.readthedocs.io/en/latest/readme.html

Mackey, T., Purushothaman, V., Li, J., Shah, N., Nali, M., Bardier, C., . . . Cuomo, R. (2020). Machine Learning to Detect Self-Reporting of Symptoms, Testing Access, and Recovery Associated With COVID-19 on Twitter: Retrospective Big Data Infoveillance Study. *JMIR public health and surveillance*, *6*(2), e19509-e19509. https://doi.org/10.2196/19509

MacKian, S. (2003). A review of health seeking behaviour: problems and prospects. *Health Systems Development Programme*.

Mamidi, R., Miller, M., Banerjee, T., Romine, W., & Sheth, A. (2019). Identifying Key Topics Bearing Negative Sentiment on Twitter: Insights Concerning the 2015-2016 Zika Epidemic. *JMIR public health and surveillance*, *5*(2), e11036-e11036. https://doi.org/10.2196/11036

Medicine, U. S. N. L. o. *National Library of Medicine - National Institutes of Health*. Retrieved March, 01 from https://www.nlm.nih.gov/

Milani, F. (2020). COVID-19 outbreak, social response, and early economic effects: a global VAR analysis of cross-country interdependencies. *Journal of population economics*, *34*(1), 223-252. https://doi.org/10.1007/s00148-020-00792-4

Moorhead, S. A., Hazlett, D. E., Harrison, L., Carroll, J. K., Irwin, A., & Hoving, C. (2013). A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *Journal of medical Internet research*, *15*(4), e85-e85. https://doi.org/10.2196/jmir.1933

Murdock, J., & Allen, C. (2015). Visualization techniques for topic model checking. Proceedings of the AAAI Conference on Artificial Intelligence,

Mykhalovskiy, E., & Weir, L. (2006). The global public health intelligence network and early warning outbreak detection. *Canadian journal of public health*, *97*(1), 42-44.

Neiger, B. L., Thackeray, R., Burton, S. H., Thackeray, C. R., & Reese, J. H. (2013). Use of twitter among local health departments: an analysis of information sharing, engagement, and action. *Journal of medical Internet research*, *15*(8), e177-e177. https://doi.org/10.2196/jmir.2775

Number of social media users worldwide 2010-2021 | Statista. (2019). https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Nussbaumer-Streit, B., Mayr, V., Dobrescu, A. I., Chapman, A., Persad, E., Klerings, I., . . . Gartlehner, G. (2020). Quarantine alone or in combination with other public health measures to control COVID-19: a rapid review. *Cochrane database of systematic reviews*, *4*, CD013574-CD013574. https://doi.org/10.1002/14651858.CD013574

O'reilly, T. (2007). What is Web 2.0: Design patterns and business models for the next generation of software. *Communications & strategies*(1), 17.

Odlum, M., & Yoon, S. (2015). What can we learn about the Ebola outbreak from tweets? *AJIC: American Journal of Infection Control*, *43*(6), 563-571. https://doi.org/10.1016/j.ajic.2015.02.023

Odlum, M., & Yoon, S. (2018). Health Information Needs and Health Seeking Behavior During the 2014-2016 Ebola Outbreak: A Twitter Content Analysis. *PLoS currents*, *10*. https://doi.org/10.1371/currents.outbreaks.fa814fb2bec36e29b718ab6af66124fa

Olenja, J. (2003). Editorial Health seeking behaviour in context. *East African medical journal*, *80*(2), 61-62.

Pallant, J. (2013). *SPSS survival manual*. McGraw-hill education (UK).

Park, H., & Park, M. (2014). Cancer Information-Seeking Behaviors and Information Needs Among Korean Americans in the Online Community. *The Publication for Health Promotion and Disease Prevention*, *39*(2), 213-220. https://doi.org/10.1007/s10900-013-9784-8

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*.

Prybutok, L. G., Koh, R. C., & Prybutok, R. V. (2014). A Content Relevance Model for Social Media Health Information. *CIN: Computers, Informatics, Nursing*, *32*(4), 189-200. https://doi.org/10.1097/CIN.0000000000000041

Qasem, M., Thulasiram, R., & Thulasiram, P. (2015). Twitter sentiment classification using machine learning techniques for stock markets. 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI),

Rees, C. E., & Bath, P. A. (2000). The psychometric properties of the Miller Behavioural Style Scale with adult daughters of women with early breast cancer: a literature review and empirical study. *Journal of Advanced Nursing*, *32*(2), 366-374.

Rees, C. E., & Bath, P. A. (2001). Information-seeking behaviors of women with breast cancer. Oncology nursing forum,

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks,

Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., & Harabagiu, S. M. (2012). EmpaTweet: Annotating and Detecting Emotions on Twitter. Lrec,

RStudio. (2019). https://rstudio.com/products/rstudio/

Saeidi, A. M., Hage, J., Khadka, R., & Jansen, S. (2015). ITMViz: Interactive topic modeling for source code analysis. 2015 IEEE 23rd International Conference on Program Comprehension,

Sesagiri Raamkumar, A., Tan, S. G., & Wee, H. L. (2020). Measuring the Outreach Efforts of Public Health Authorities and the Public Response on Facebook During the COVID-19 Pandemic in Early 2020: Cross-Country Comparison. *Journal of medical Internet research*, *22*(5), e19334. https://doi.org/10.2196/19334

Shah, K., Kamrai, D., Mekala, H., Mann, B., Desai, K., & Patel, R. S. (2020). Focus on Mental Health During the Coronavirus (COVID-19) Pandemic: Applying Learnings from the Past Outbreaks. *Curēus (Palo Alto, CA)*, *12*(3), e7405-e7405. https://doi.org/10.7759/cureus.7405

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, *5*(4), 13-22.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. Proceedings of the workshop on interactive language learning, visualization, and interfaces,

Skinner, H., Biscope, S., & Poland, B. (2003). Quality of internet access: barrier behind internet use statistics. *Social Science & Medicine*, *57*(5), 875-880.

Sloan, L., & Quan-Haase, A. (2017). *The SAGE handbook of social media research methods*. Los Angeles ; London : SAGE reference.

Smailhodzic, E., Hooijsma, W., Boonstra, A., & Langley, D. J. (2016). Social media use in healthcare: A systematic review of effects on patients and on their relationship with healthcare professionals. *BMC Health Services Research*, *16*(1). https://doi.org/10.1186/s12913-016-1691-0

Soares, A. M., Pinho, J. C., & Nobre, H. (2012). From social to marketing interactions: The role of social networks. *Journal of Transnational Management*, *17*(1), 45-62.

Sokolova, M., Jafer, Y., & Schramm, D. (2012). Text Mining for Personal Health Information on Twitter. In (pp. 112-112).

Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2014). Social Media Analytics: Ein interdisziplinärer Ansatz und seine Implikationen für die Wirtschaftsinformatik: An

Interdisciplinary Approach and Its Implications for Information Systems. *WIRTSCHAFTSINFORMATIK*, *6*(2), 101-109. https://doi.org/10.1007/s12599-014-0315-7

Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2018). Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter. *Political communication*, *35*(1), 50-74. https://doi.org/10.1080/10584609.2017.1334728

Teufel-Shone, N. I., Cordova-Marks, F., Susanyatame, G., Teufel-Shone, L., & Irwin, S. L. (2015). Documenting Cancer Information Seeking Behavior and Risk Perception in the Hualapai Indian Community to Inform a Community Health Program. *Journal of community health*, *40*(5), 891-898. https://doi.org/10.1007/s10900-015-0009-1

The Twitter Rules. (2019, 2019-11-01 12:59:49). Twitter. https://help.twitter.com/en/rules-and-policies/twitter-rules

Thomas, L. G., & Mark, S. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences - PNAS*, *101*(Suppl 1), 5228-5235. https://doi.org/10.1073/pnas.0307752101

Tipping, G. (1995). *Health care seeking behaviour in developing countries : an annotated bibliography and literature review / by Gill Tipping, Malcolm Segall*. Brighton : Institute of Development Studies at the University of Sussex.

Twitter Usage Statistics - Internet Live Stats. (2019). https://www.internetlivestats.com/twitter-statistics/

Underhill, C., & McKeown, L. (2008). Getting a second opinion: health information and the Internet. *Health Reports*, *19*(1), 65.

United Nations. *Coronavirus global health emergency: Coverage from UN News | | UN News*. Retrieved January, 22 from https://news.un.org/en/events/coronavirus-global-health-emergency-coverage-un-news

United Nations. (2019). *Coronavirus global health emergency: Coverage from UN News | | UN News*. Retrieved January, 22 from https://news.un.org/en/events/coronavirus-global-health-emergency-coverage-un-news

Valli, P. A., Uma, M., & Sasikala, T. (2017). Tracing out various diseases by analyzing Twitter data applying data mining techniques. 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS),

Vaus, D. d. (2002). Analyzing social science data. In *California: Sage Publication*.

Vega, E., Parthasarathy, R., & Torres, J. (2010). Where are my tweeps?: Twitter usage at conferences. *Paper, Personal Information*, 1-6.

Waheed, H., Anjum, M., Rehman, M., & Khawaja, A. (2017). Investigation of user behavior on social networking sites.(Research Article). *PLoS ONE*, *12*(2), e0169693. https://doi.org/10.1371/journal.pone.0169693

Wang, J., Parsey, C., Davis, B., Cheng, T. Y.-m., Liu, L., & Woo, B. K. P. (2018). Analyzing Twitter as a Platform for Alzheimer-Related Dementia Awareness: Thematic Analyses of Tweets. *JMIR Aging*, *1*(2). https://doi.org/10.2196/11542

Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in NLP. COLING 1992 Volume 4: The 15th International Conference on Computational Linguistics,

WHO. (2020a). *WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020*. Retrieved 18, May from https://www.who.int/director-

general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

WHO. (2020b). *World Health Organization*. World Health Organization. https://www.who.int/news/item/29-06-2020-covidtimeline

WHO. (2021). *WHO Coronavirus (COVID-19) Dashboard*. https://covid19.who.int/

Wikipedia. (2019). *Twitter --- {Wikipedia}, The Free Encyclopedia*. https://en.wikipedia.org/wiki/Twitter

Williams, M. L., Burnap, P., & Sloan, L. (2017). Crime Sensing With Big Data: The Affordances and Limitations of Using Open-source Communications to Estimate Crime Patterns. *British Journal Of Criminology*, *57*(2), 320-340. https://doi.org/10.1093/bjc/azw031

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *<h>Data</h> <h>Mining</h>*. San Francisco: Elsevier Science & Technology.

Yonker, L. M., Zan, S., Scirica, C. V., Jethwani, K., & Kinane, T. B. (2015). "Friending" teens: systematic review of social media in adolescent and young adult health care. *Journal of medical Internet research*, *17*(1), e4.

Zhang, H., Lu, Y., Gupta, S., & Zhao, L. (2014). What motivates customers to participate in social commerce? The impact of technological environments and virtual customer experiences. *Information & Management*, *51*(8), 1017-1030. https://doi.org/10.1016/j.im.2014.07.005

Zhang, J., Chen, Y., Zhao, Y., Wolfram, D., & Ma, F. (2019). Public health and social media: A study of Zika virus-related posts on Yahoo! Answers. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24245

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011*, *89*.

Zimmer, M., & Hoffman, A. (2012). 16 Privacy, Context, and Oversharing: Reputational Challenges in a Web 2.0 World. *The reputation society: How online opinions are reshaping the offline world*, 175.