UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

A FRAMEWORK OF EPIDEMIC MODELING CONSIDERING PREVENTIVE BEHAVIORS: COMPARTMENTAL MODELING, TEXT ANALYTICS, AND MACHINE LEARNING

A Dissertation

Submitted to the Graduate Faculty

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

By

HYEJIN CHO

Norman, Oklahoma

2024

A FRAMEWORK OF EPIDEMIC MODELING CONSIDERING PREVENTIVE BEHAVIORS: COMPARTMENTAL MODELING, TEXT ANALYTICS, AND MACHINE LEARNING

A DISSERTATION APPROVED FOR THE
INDUSTRIAL AND SYSTEMS ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Randa Shehab, Chair

Dr. Charles Nicholson

Dr. Shivakumar Raman

Dr. Rui Zhu

Dr. Dean Hougen

Dr. Hairong Song

© Copyright by HYEJIN CHO 2024
All Rights Reserved.

A special feeling of gratitude goes to my loving mom and dad, Hyun J and Kye S, for being there whenever I found myself lost in the wilderness during my doctoral journey. A special thanks to my brother, Daniel Tae, my sister-in-law, and my nephew, for standing by me and encouraging me throughout my doctoral studies. I also dedicate this dissertation to my grandmother, Gemma Y, who ascended to heaven while I pursued my studies. I also thank all my family members, my grandmother and grandfather, my aunties and uncles, and many friends for cheering me on to complete this long marathon. Lastly, I dedicate my dissertation to the glory of God.

ACKNOWLEDGMENTS

Grateful Acknowledgement is made to the School of Industrial and Systems Engineering at the University of Oklahoma for their very generous support of my doctoral study, and to the Data Science and Analytics Institute and the College of Engineering for providing exceptional resources and space to research and experiment.

I am equally indebted to friends and family. Special thanks to the wonderful staff, Melodi Franklin, so her fellow staff Jennifer Ille, and to my old friend, Cheryl Carney, everyone supportive. Thanks to Director Henry Neeman of the OU Supercomputing Center for helping me to utilize OSCER; to my beautiful mentor, Dr. Kimberly Wolfinbarger for her warm and encouraging lessons for teamwork and leadership; to Dr. Andres Gonzalez for allowing me to assist ISE5023; to Dr. Doyle Dodd, such a humorous and supportive mentor, for allowing me to assist ISE4804; to Dr. Theodore Trafalis, for sharing me with good books for optimization when we all evacuated to the Crossroads at the banquet; to Dr. Kash Barker, for supporting my doctoral study and organizing important seminars; to Dr. Christan Grant, for instructing me in CS5293 and answering all my questions throughout the course; to Dr. Talayeh Razzaghi, for providing me several references for early pieces of my study; to Dr. Ziho Kang, such a smart and caring advisor, for allowing me to assist ISE5663 and ISE5853; to Dr. Rui Zhu and Dr. Hairong Song, for agreeing to serve on my committee; to my committee, Dr. Charles Nicholson, who deftly edited several mathematical problems in my study; and to my committee, Director Dean Hougen, who cleverly guided me through the modeling processes for compartmental differential equations and helped me improve computational experiments; and especially to my wonderful supervisor, Dean Randa Shehab, such a brilliant and insightful advisor and life-changing mentor, who has believed in my study from the beginning and has guided me to shape my entire doctoral study to pursue my academic vision. It is to her that I have dedicated this dissertation, with grati-

tude. Finally, I would like to acknowledge my department head, Director Shivakumar Raman, who, after hearing my vision of pursuing Artificial Intelligence and Production and Healthcare Systems, looked me in the eye one day and introduced me to Dean Shehab.

TABLE OF CONTENTS

LIST OF TABLES xi

LIST OF FIGURES xiii

ABSTRACT xv

1 INTRODUCTION 1

1.1 Research Questions 2

1.2 Scope 3

1.3 Background Information 3

1.3.1 COVID-19 3

1.3.2 Social networks 9

2 LITERATURE REVIEW 15

2.1 Epidemiological Models 15

2.1.1 Markov process 16

2.1.2 Compartmental modeling 17

2.1.3 Parameter estimation 23

2.2 Artificial Intelligence 26

2.2.1 Machine learning 27

2.2.2 Support vector machine 28

2.2.3 K-nearest neighbor 30

2.2.4 Semi-supervised learning 31

2.3 Text Analytics 32

2.3.1 Text classification 33

2.3.2 Topic modeling 33

2.3.3 Sarcasm detection 37

2.3.4 Bidirectional encoder representations from Transformers 39
2.3.5 Sentiment analysis 39

3 DATA 43
3.1 Scoping 43
3.2 Description of the Data 44
3.2.1 COVID-19 cases 46
3.2.2 Twitter (X) 46
3.3 Process for Retrieving the Data 48
3.3.1 Planning 48
3.3.2 Processing 49

4 PREDICTION OF INDIVIDUAL PREVENTION BEHAVIORS USING SOCIAL MEDIA DATA 52
4.1 Problem Description 52
4.2 Data 52
4.2.1 Data preprocessing 54
4.2.2 Encoding (prevention behaviors lexicon) 57
4.3 Prediction of Individual Prevention Behaviors 59
4.4 Results and Discussion 61
4.4.1 Tweets results 61
4.4.2 Training results 63
Number of iteration of self-training 63
Hyper-parameters 63
4.4.3 Model evaluation results 63
Baseline vs support vector machines 63
K-nearest neighbor 64
4.4.4 Prediction results 66

Individual tweets . . . 66
Aggregated prediction results . . . 66
4.4.5 Discussion of the Prediction of Individual Prevention Behaviors . 67

5 INCORPORATION OF THE EFFECT OF PREVENTION BEHAVIORS WITH EPIDEMIC MODELING . . . 71
5.1 Problem Formulation . . . 71
5.2 Data & Experiment Settings . . . 74
5.2.1 Data . . . 74
5.2.2 Public policies - interventions . . . 74
5.2.3 Estimation of transmission rate - deterministic dynamic variable . 75
5.3 Methods . . . 77
5.3.1 Grid search method . . . 77
5.3.2 Trust region method . . . 78
Barrier function in trust region interior point . . . 78
Convergence of trust region interior point . . . 79
Experiment setting . . . 80
5.3.3 Alternating minimization . . . 81
5.3.4 Basic reproduction number of COVID-19 . . . 82
5.3.5 Model validation - bootstrap re-sampling . . . 82
5.4 Result . . . 82
5.4.1 Parameter estimates . . . 82
5.4.2 Model validation . . . 85
5.5 Discussion . . . 86
5.5.1 Parameter estimates . . . 86
5.5.2 Comparison between grid search and trust region . . . 87
5.5.3 Performance of grid search . . . 87

5.5.4 Nonlinear constrained optimization to unconstrained optimization in the trust region interior point method . . . 88
5.5.5 Trust region and initial starting point . . . 88
5.5.6 Aggregated effect of recovered rate and deceased rate in the SEIRD 89
5.5.7 Subsequent alternating minimization of the mortality rate . . . 90
5.5.8 Measure of the effects of nonpharmaceutical interventions . . . 90

6 CONCLUSION . . . 92
6.1 Evaluation of the Compliance with Individual Prevention Behaviors . . . 92
6.2 Incorporation of the Effect of Preventive Behaviors with Epidemic Modeling . . . 94
6.3 Contribution . . . 95

7 LIMITATION & FUTURE WORK . . . 97

APPENDICES . . . 115
A1 COVID-19 Cases . . . 115
A2 Pseudo Code: PRecomm . . . 130

LIST OF TABLES

1.1 COVID-19 Timeline (CDC, 2023a) 4

1.2 Coronavirus variants with dates and locations when first detected (WHO (2023), CDC (2023)) 7

3.1 Starting and ending dates of stay-home orders by state (as of May 31, 2020) 44

3.2 United States COVID-19 Cases and Deaths (%) by State (as of May 31, 2020) 45

3.3 Data specification of the collected COVID-19 cases 45

3.4 Data specification of the collected tweets 46

3.5 The keywords used in searching tweets between March 1, 2020 and May 31, 2020 (Chen et al. (2020a)) 47

3.6 Search terms for the state of New York used in the python codes 48

3.7 The number of tweets (%) in the inclusion and the exclusion criteria (2020) . 50

3.8 Example of the collected tweets 51

4.1 Examples of the collected tweets (Tweet ID's and user names are hidden) . . 55

4.2 Interventions (prevention behaviors) by CDC 57

4.3 The number of tweets of labeled tweets, labeled tweets after balancing, unlabeled tweets, and total tweets in the state of New York 61

4.4 The number of tweets in each prevention behavior after balancing (the number of tweets before balancing is described in parenthesis) 62

4.5 Model performance metrics of random forest (RF), support vector machines (SVM), k-nearest neighbor (KNN) 65

4.6 An example of the PRecomm outcome 70

5.1 Notation 73

5.2 COVID-19 Timeline in New York 75

5.3 Range of parameter values for grid search 77

5.4 Comparison of the mean squared errors, parameter estimates, and computation times among sequential least squares, linear approximation, and trust region interior point 80

5.5 Experiment setup for the trust-region interior-point method 81

5.6 Parameter estimates by grid search and trust region in this study compared to the values from a baseline study (Chowell et al. (2003)) 84

LIST OF FIGURES

1.1 The number of infected cases by state over 2020-2022 6

1.2 Increase in social media use, 2020 10

1.3 Triadic reciprocal causation 13

2.1 Conceptual overview of text classification 34

2.2 LDA and decomposition of document-word matrix into document-topic matrix and topic-word matrix (Seth, 2021) 35

2.3 Algorithm - Latent Dirichlet Allocation (LDA) 36

2.4 Pattern extraction 37

4.1 A framework of disaster management systems with prediction of individual prevention behaviors 53

4.2 Inclusion and exclusion criteria of tweets retrieval 54

4.3 Word cloud of collected tweets in the state of New York 56

4.4 Prediction and Recommended Predictions (PRecomm) 60

4.5 Original (a) and balanced (b) prevention behaviors distribution 62

4.6 F-1 scores for hyper-parameters tuning (a) random forest, (b) support vector machines, and (c) k-nearest neighbor in the initial training and each iteration of self-training 64

4.7 F-1 scores and prediction probabilities or similarity scores by random forest, support vector machines, and k-nearest neighbor 65

4.8 Prediction scores distribution - SVM 66

4.9 Prediction scores distribution without (a) and with self-training (b) 67

4.10 Confusion matrix at each iteration of self-training with random forest (a ratio of training data to pseudo-labeled data was 90%:10% at each iteration) . . . 68

5.1 The SEIRD model with nonpharmaceutical interventions and model parameters 71

5.2 A theoretical step function of the dynamic variable of transmission rate assuming a public policy is implemented on March 23 and another public policy is implemented on April 30 76

5.3 MSE by SLSQP, COBYLA, Trust region interior point, and a combined graph of results from all three methods 79

5.4 Trust region results - (a) the consequent mean squares of errors (b) parameter estimates with different initial starting points 85

5.5 Trust region and alternating optimization results – (a) Beta distribution and (b) prediction of coronavirus dynamics using SEIRD model with the optimal parameters from trust region and alternating optimization 85

5.6 Grid search and Trust region 87

5.7 Reduced error in the deceased cases – (a) Before alternating optimization (b) after alternating optimization 90

7.1 Heterogeneous SEIRD with variables and weights 98

ABSTRACT

Cho, Hyejin. Ph.D., University of Oklahoma, May 2024, A Framework of Epidemic Modeling Considering Preventive Behaviors: Compartmental Modeling, Text Analytics, and Machine Learning. Major Professor: Dr. Randa Shehab

The goal of this research is to provide a novel framework for epidemic modeling incorporating metrics derived from social media to predict epidemic dynamics and to estimate the impact of preventive behaviors. This study employs empirical data collected from Centers for Diseases Control and Prevention, and Twitter (or X) to demonstrate the practical usability of the proposed framework. Specifically, this research utilizes optimization, simulation, and compartmental differential equations to predict the number of infected and deceased individuals. The research estimates the basic reproduction number (R_0) for diseases dynamics. In addition, this study utilizes artificial intelligence and develops a self-training machine learning algorithm to predict the individual compliance level with prevention behaviors. In the analysis, the effect of preventive behaviors on mitigating transmission is evaluated quantitatively. The research contributes to enhance the accuracy of epidemic modeling and to improve decision-making within public healthcare systems, ultimately leading to a reduction in mortality rates and the saving of more lives.

1. INTRODUCTION

Infectious diseases are one of the serious disasters particularly in developing countries. The Ebola outbreak in West Africa in the 2014-2016 was recorded as the most significant Ebola epidemic that has occurred worldwide since the virus was first discovered (Kaner and Schaack (2016)). About 40% of people infected with Ebola died (Cho, 2016). In the subsequent years, there was a significant concern regarding the outbreak of the Middle East Respiratory Syndrome coronavirus (MERS-CoV), particularly notable in Saudi Arabia from 2014 to 2016. MERS-CoV is a respiratory virus causing severe illness, with a historical fatality rate of around 35% (Donnelly et al. (2019)). Most recently, the COVID-19 outbreak was declared as a pandemic on March 10, 2020 by the World Health Organization (WHO). The pandemic resulted in the loss of numerous lives and paralyzed normal functioning of societies in both developing and non-developing countries. The epidemic has continued for more than three years since 2020, officially declared ended on May 11, 2023 (CDC, 2023b).

Nonpharmaceutical intervention was recommended by WHO and Centers for Disease Control and Prevention (CDC) even before and after coronavirus variants emerged (CDC, 2021). While vaccination played a pivotal role in mitigating the spread of the virus, policymakers implemented non-pharmaceutical interventions (NPIs) both prior to and along with the development of vaccines, particularly in the initial stage of COVID-19 transmission (Lee et al. (2020)). These interventions were necessary to reduce transmission rates and minimize the risk of mortality (Valladares et al. (2022)).

Since summer 2021, the Delta variant became a predominant virus, discovered in India, causing several surges of COVID-19 cases worldwide (Katella et al., 2023). Subsequently, Omicron and its subvariants have been a prevalent strain in the United States of America since late 2021 (Katella et al., 2023). These strains have heightened the transmission rate and diminished the effectiveness of vaccines, prolonging the COVID-

19 outbreak. Consequently, it is important that public policies aimed at alleviating person-to-person transmission be implemented aligned with vaccination efforts to effectively mitigate the spread of the virus and prevent fatalities (Antonelli et al., 2022). For instance, governments and healthcare institutions collaborated to decrease the rate of person-to-person contact by restricting individual physical movement. These strategies included social distancing, quarantine protocols, closures or lock-downs, as well as travel bans. Additionally, efforts were made to minimize the likelihood of respiratory infection through the use of masks and increased disinfection practices (CDC (2021)).

Policy-making during epidemics should include analysis of ongoing epidemic dynamics and predictions, as well as evaluating the effectiveness of proposed strategies. However, previous studies rarely perform a comprehensive estimation of coronavirus dynamics in evaluating the impacts of non-pharmaceutical interventions (NPIs). In many cases, either parameter estimation or evaluation of NPI effects is missing. For example, studies that estimate the effects of NPIs often adopted disease dynamics such as transmission rates from past studies without considering regional or temporal differences, population density, or contact rates (Enns et al., 2020; Carcione et al., 2020; Yarsky, 2021).

1.1 Research Questions

It is important to predict epidemic dynamics and to research the effects of mitigation strategies for reducing transmission and preventing mortality. Motivated by this perspective, I posed following research questions: (i) Can the accuracy of epidemic modeling be improved by incorporating the effect of preventive strategies?; (ii) how to quantify the effect of mitigation strategies?; and (iii) how to predict people's adoption of preventive behaviors during an epidemic?. The sequence of questions reflects the thought process behind the proposed research. The second question can be partially answered by the third question. Thus, to address the initial two questions, the third question should be addressed as a priority. To sum up, the research questions tackled in this

study encompasses: (i) how to predict the extent of individual adoption of preventive behaviors during an epidemic by using social media data and (ii) can we improve the accuracy of epidemic modeling by incorporating the effect of preventive behaviors on transmission. The outcomes from this research can be utilized in policy making process for disaster management.

1.2 Scope

The primary focus of this research is to construct a framework for epidemic modeling that integrates the measured impact of preventive behaviors to enhance the accuracy of predicting disease dynamics. The study analyzed the effects of preventive behaviors using data extracted from social media. The discussion includes the derived estimates of preventive behaviors obtained through epidemic modeling, coupled with an analysis of social media data. Consequently, this study considers pertinent research on epidemic modeling, analyzes COVID-19 data for demonstration of epidemic modeling considering preventive behaviors, and the utilization of social media in epidemic contexts to predict individual preventive behaviors.

1.3 Background Information

1.3.1 COVID-19

The very first cases of COVID-19 were reported in Wuhan China on December 1, 2019. A group of individuals with symptoms of unknown cause were reported (Huang et al., 2020). The World Health Organization (WHO) was informed of the cases with pneumonia-like symptoms from Wuhan's Hunan Seafood Wholesale Market on 31 December, 2019, resulting in the shut down of the seafood market on January 1 2020. Huang et al. (2020) identified the unknown disease as a novel coronavirus, similar to the one associated with SARS and the MERS. New cases were reported in neighboring

countries, including Japan and Korea as early as Janunary 20, 2020 (Shim et al., 2020). The first case in the United States was confirmed on January 20, 2020 (CDC, 2023a). A surge in transmission in Italy was reported February 23, 2020, resulting in a national shutdown (Cavallaro et al., 2021). On March 11, 2020, the WHO declared COVID-19 a pandemic. There were more than 118,000 cases in 114 countries and 4,291 deaths (CDC, 2023a). Table 1.1 describes the major events during the COVID-19 outbreak. This information provides the starting time with location of the disease, efforts made to mitigate the disease, public concerns, etc.

Table 1.1. COVID-19 Timeline (CDC, 2023a)

Date	Year	Location	Description
December 1	2019	Wuhan, China	The first symptoms for COVID-19
March 11	2020	Worldwide	Declaration of COVID-19 pandemic by the WHO
March 13	2020	U.S.A	Declaration of Nationwide emergency and travel ban by the Trump Administration
March 28	2020	U.S.A	Implementation of social distancing, quarantine for 14 days
April 3	2020	U.S.A	Implementation of mask wearing
April 4	2020	Worldwide	More than 1 million cases of COVID-19 has been confirmed

continued on next page

Table 1.1. *continued*

Date	Year	Location	Description
May 26	2020	U.S.A	Implementation of lockdowns, curfews, stay-at-home orders, masking, checkpoints by Navajo officials
May 28	2020	U.S.A	The recorded death toll from COVID of more than 100,000
December 11	2020	U.S.A	Recommendation of Pfizer's COVID-19 vaccine for all people ages 16 years or older
December 19	2020	U.S.A	Recommendation of the Moderna COVID-19 vaccine in persons ages 18 years or older
January 18	2021	U.S.A	The reported death toll from COVID-19 of more than 400,000

Novel coronavirus (SARS-CoV-2) variants has threatened immunity for COVID-19, increasing transmissibility and extending infection duration (Van Egeren et al., 2021). Figure 1.1 presents the number of infected cases over a two-year period in various states where the number of infected cases were the most serious. WHO defined variants as viruses evolve as they spread among people over time then these changes become different from the original virus, which are known as variants (Burki, 2021). During the COVID-

19 outbreak, coronavirus variants were named by the Greek alphabet such as alpha, beta, gamma, delta, or omicron (nu) in May of 2021 (WHO, 2021). Table 1.2 describes the timeline of variants. Although this study analyzes coronavirus dynamics considering only original coronavirus, information regarding the variants is included for future study.

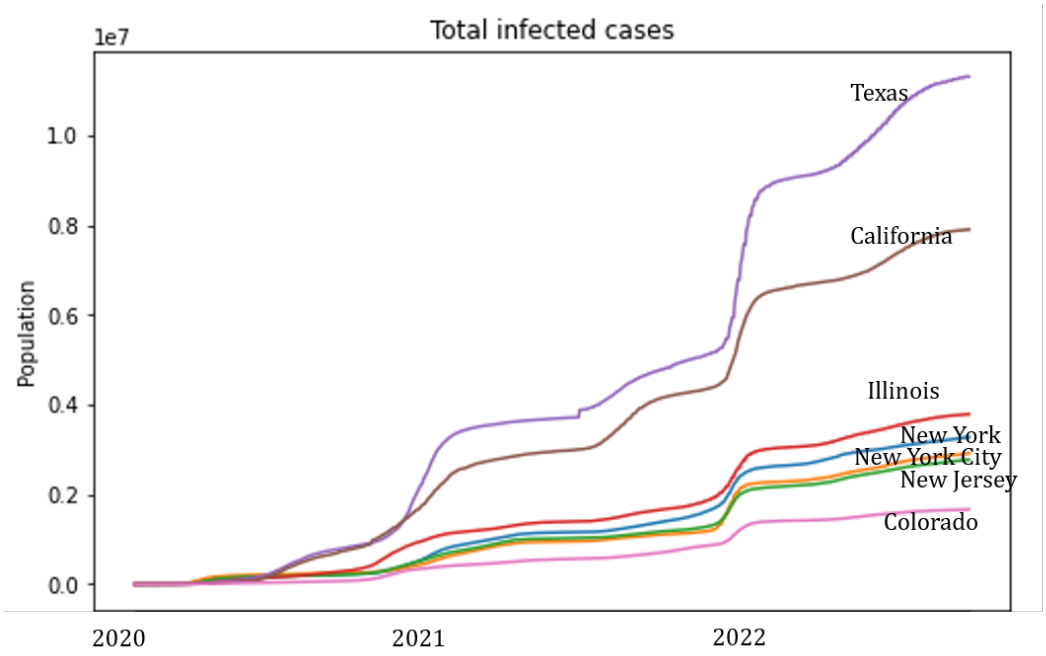


Figure 1.1. The number of infected cases by state over 2020-2022

Nonpharmaceutical interventions (NPIs) in response to infectious diseases have been only alternative available to mitigate the spread of diseases before the development of vaccines. Particularly, the importance of NPIs is heightened in developing countries where access to medical services and medicines are incapacitated. During the COVID-19 outbreak, populations have experienced delays in vaccination and medication even in countries with well-established healthcare systems (Wilder-Smith and Freedman, 2020). This disorder has highlighted the importance of NPIs in combination with medical treatment for reducing the spread of diseases, which underscores the need for quantifying the impact of NPIs for policymakers to effectively address epidemics.

Table 1.2. Coronavirus variants with dates and locations when first detected (WHO (2023), CDC (2023))

Variants	First case	First case (USA)	Note
Alpha	September 2020, UK	29 December 2020, Colorado 22 January 2021, 12 states	– Universal mitigation strategies (CDC)
Beta	May 2021, South Africa	28 January 2021, South Carolina	–
Gamma	November 2020, Brazil	25 January 2021, Minnesota	–
Delta	October 2020, India	1 June 2021	Third wave of infections
Omicron	November 2021, Botswana, South Africa	1 December 2021, California, San Francisco 2 December 2021, Minnesota, New York City	More than ten times infectious than the delta wave A second case in the U.S.

Various mitigation strategies aimed at reducing disease transmission categorized as mask, hygiene, social distancing, and quarantine or isolation, have been adopted for public health purposes during the COVID-19 pandemic (CDC, 2021; Guy et al., 2021). Haug et al. (2020) provided the detail of NPIs, for instance, social distancing was specified by small gathering cancellation, mass gathering cancellation, closure of educational institutions, measures for special populations, etc.

Among public policies during the COVID-19 pandemic, limiting physical distances between individuals such as social distancing and quarantine has been acknowledged as an effective policy in mitigating the transmission of epidemics (Shin, 2021; Chowell et al., 2003). Tsay et al. (2020) emphasized that quarantine was the most crucial strategy, fol-

lowed by social distancing and lockdown interventions during the COVID-19 pandemic. Screening and testing for the disease were found to be vital, particularly prior to periods of relaxed social distancing (Tsay et al., 2020). Liu et al. (2021) reported travel restrictions, quarantine, and distancing are potentially effective in delaying COVID-19 spread. In Haug et al. (2020), the most effective NPIs included curfews, lockdowns and restricting small or large gathering, while individual movement restrictions were also one of the top-ranked interventions. Some studies explored the impact of intervention strategies on COVID-19 and other infection such as HIV/AIDS, which has been prevalent in a developing country. For instance, Teklu and Kotola (2023) investigated the impact of intervention strategies on HIV/AIDS and COVID-19 co-infection transmission, showing increasing treatment intervention highly decreases the number of co-infectious population, emphasizing the efficacy of intervention.

Several studies acknowledged the significance of quantifying these mitigation policies during COVID-19. However, although some studies have attempted to quantify the efficacy of NPIs, they often derived their estimates by assuming disease characteristics from published literature that had different study scopes compared to their own research. Tang et al. (2020) added compartments of quarantine and isolation to the classical SEIR model to quantify the effectiveness of quarantine and isolation in Wuhan, Hubei, China using the Markov Chain Monte Carlo (MCMC) method. Chinazzi et al. (2020) studied the effect of travel ban in Wuhan using a SEIR model and the global epidemic and mobility model. A metapopulation network in Chinazzi et al. (2020) comprises sub-populations connected by mobile individuals, with data obtained from the population database at the Socioeconomic Data and Application Center at Columbia University. Parameters in Chinazzi et al. (2020) including the latent and infectious period, and generation time were derived from previous publications. In Enns et al. (2020), COVID-19 spread in Minnesota was modeled by adding 4 additional compartments (i.e., subclinical infection, symptomatic infection, hospitalized and not ventilated, and ICU

and ventilated) to the SEIRD model. In the modeling process in Enns et al. (2020), the incubation and infectious periods were derived from a study conducted in Wuhan, while the range of basic reproduction number was derived from a study conducted in European countries. Then, the transmission rate was adjusted to be consistent with the basic reproduction number (Enns et al., 2020). Additionally, social distancing was assumed to reduce contacts by 50%, while shelter in place was assumed to reduce contacts by 80% in Enns et al. (2020).

Therefore, it has been investigated that while several studies have considered the significance of quantifying the effectiveness of NPIs, they have estimated the impact of NPIs based on assumptions used in previously published studies that differed in terms of time, location, and population density. Hence, it is unavoidable that the analysis might not fully reflect the true values, since the study did not use the same data or adjust it according to their specific research focus. To address this limitation, our study involves estimating the parameters of COVID-19 evolution inspired by empirical data and subsequently evaluating the impact of NPIs based on these estimated parameters.

1.3.2 Social networks

Statistics in Clement (2020) show that, in 2020 approximately 3.6 billion people were using social media (e.g., YouTube, Facebook, and Twitter) worldwide, and the number was projected to continue to grow. A survey study examined the impact of the pandemic also showed that increase on social media use for the United States adults in Figure 1.2 (Insider Intelligence, 2020). About 5-25% of respondents answered they increased their usage on social media, with YouTube showing the largest increase and Twitter showing about 10% increase during the COVID-19 outbreak in May 2020. With such a large membership and growing usage, social media has become one of the popular methods to engage in dialogue and exchange information in emergency situations (Liu, 2021). One of the striking features of social media is the relationship between human behaviors in social

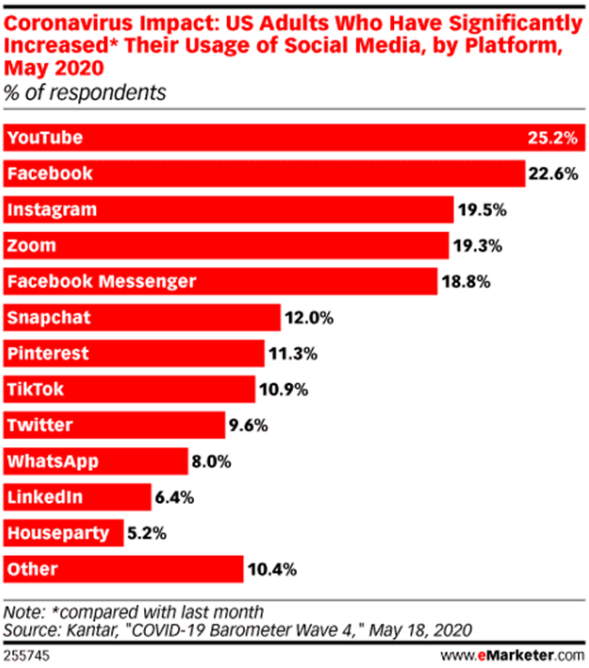


Figure 1.2. Increase in social media use, 2020

media, which is bidirectional and interactive (Liu, 2021). In other words, exposure to certain information or messages was associated with individual cognition, which in turn, influenced their behaviors (Bandura, 1984). Hu et al. (2018) found a similar pattern such that user engagement in online communities corresponded to activities in physical world.

Government agencies also leverage social media to encourage citizen engagement in disaster management. For instance, during the 2012 Hurricane Sandy crisis in the United States, government agencies adopted Twitter to engage citizens in critical public services development (Chen et al., 2020b) such as digital volunteer organizations (Meier, 2010). Another instance is the United Nations Offices for the Coordination of Humanitarian Affairs. They collaborate with online communities of volunteer organizations and extract information from social media for responding emergency situation (Meier, 2010).

However, data credibility is one of the concerns about social media data. Cinelli et al. (2020) analyzed narratives and moods in five different social media platforms (e.g., Twitter, Instagram, YouTube, Reddit and Gab) during the COVID-19 outbreak. Cinelli et al. (2020) first clustered news into two groups: reliable or questionable sources based on guidelines shared by online fact checking organizations using Partitioning Around Medoids (PAM) algorithm with cosine distance. Then, the authors evaluated the ratio of reliable to questionable sources across platforms and analyzed the number of newly created posts relevant to coronavirus across social media. Lastly, the authors modeled the spread of information with epidemic models and evaluated the basic reproduction number of the mis- or disinformation. The study included more than 8 million comments and posts related to coronavirus over 45 days (1st of January to the 14th of February) from worldwide services. The authors used the stochastic gradient descent with back-propagation rule for contents representation. The study found that users in Twitter, Instagram, YouTube were less susceptible to diffusion of information from questionable sources. Also, it found that information deriving from news marked as reliable or questionable did not present a significant difference in the way it spreads. The study showed that among mainstream social media, Twitter was the most neutral, whereas YouTube amplifies questionable sources less. Cinelli et al. (2020) concluded that the main drivers of information spreading are related to specific peculiarities of platform.

Bae et al. (2021) studied the effect of misinformation on social media in epidemic modeling. Misinformation include information designed to harass specific targets or inaccurate information or fake news on purpose of deceiving individuals. Bae et al. (2021) assumed that those negative functionalities of social media were associated with an increase in the transmission rate of the epidemic and modeled it by a penalty term in epidemic model. Chan et al. (2021) studied public opinions that were malignant to individual motivation. For instance, dissemination of negative opinions of face masks

evokes discrimination toward and labeling of individuals who adhere to those advised preventive behaviors (Chan et al., 2021).

Furthermore, individuals and organizations need to know what is happening with the disaster response and determine how they might help. Social media can facilitate this process (Sharf and Rahman, 2018). Furthermore, social media can be particularly helpful in connecting victims, survivors from disaster, and susceptible people to other individuals. It facilitates attitudes and feelings of support, encouragement, and connection, which improve mental health. Merchant and Lurie (2020) emphasized the importance of trustworthy data on social media. While acknowledging its positive impact, particularly in times of crisis with limited resources, it is crucial to ensure data reliability when understanding user preferences for new business models. On the other hand, during crises, it is vital to recognize that individuals turn to social media as major means to connect with their family and friends, share their situation, and seek support.

Social Learning Theory

Bandura (1984) proposes that individuals are influenced by observing others' behavior, emphasizing the relationship between social media usage and COVID-19 transmission. Applying social learning theory, Bae et al. (2021) utilized Uses and Gratification theory to categorize motivations for social media engagement into social and informational motives. Specifically, the impact of social media on healthcare practices and disease prevention can be analyzed in terms of both positive and negative contributions. Regarding positive contributions, the study identified and adapted a range of functions to enhance social media's effectiveness during crises, including disseminating intervention guidelines (such as self-quarantine and social distancing), sharing information on therapies, maintaining ongoing communication with affected individuals, promoting positive behavioral changes, facilitating public relations efforts, serving as a contact point for service provision, supporting recruitment activities, gauging public opinion, enabling

online medical consultations, gathering data, spreading positive news, and potentially correcting behavior.

Information available on social media can be used to measure the spread of disease since keywords and emotional status on social media were highly correlated with the degree of the disaster in affected regions. In contrast, negative functions of social media also exist. For instance, users find a target to harass online or share inaccurate information or fake news on purpose to deceive individuals. In this study, social media use with a negative function is associated with an increase in the transmission rate of epidemic. This can be modeled by a penalty term in epidemic model.

Social Cognitive Theory

Social cognitive theory emphasizes that reciprocal causation of individual behaviors between personal factors, behavioral factors, and social environmental factors, which is called triadic reciprocal causation (Bandura, 1984; Li et al., 2020). This triadic reciprocal causation is illustrated in Figure 1.3. The distinctive aspect of social cognitive theory is that it holds human behavior is shaped and controlled by both personal cognitive (e.g., expectations, beliefs) and social network (e.g., social systems). For instance, personal factors include values, self-efficacy, outcome expectations, and behavioral factors encompass include prior behavior, and social environmental factors encompass others' behaviors and feedback (Li et al., 2020).

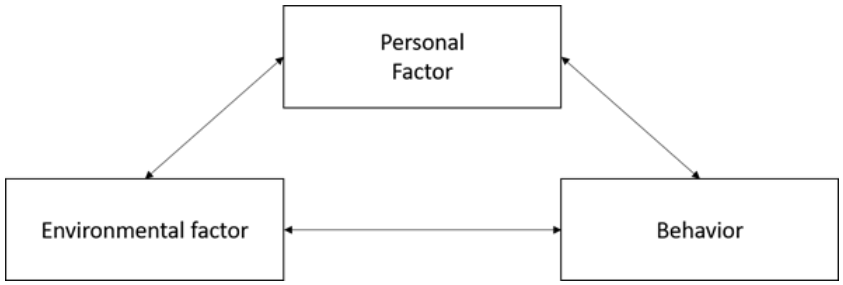


Figure 1.3. Triadic reciprocal causation

Therefore, it is important to understand the possible impact of exposure to social media to access health information or to simply observe social environmental factors on individual's preventive behavior change in a pandemic.

Social cognitive theory supports the outcome expectancy because people are motivated to perform a particular behavior if they feel driven, while self-efficacy deals with judgements of one's learning and performing actions when handling the prospective situation. In case of individual preventive behavior during COVID-19, people would be more motivated to overcome the pandemic threat by following prevention guidelines by learning about the social norm online community while social distancing (Chan et al., 2021). Regarding consumer behavior during the pandemic, it was shown that emotional states including vulnerability and pressure affect consumption decision making, which consequently influences their personal motivation and behavior process (Kursan Milaković, 2021). Also, several studies advocate that self-efficacy is an important driver of behavioral processes (Chan et al., 2021; Kursan Milaković, 2021).

Social cognitive theory has been applied to explore health promotion and disease prevention, based on the belief that individuals possess significant control over their health. Moeini et al. (2019) investigated how alterations in constructs of Social Cognitive Theory correlate positively with changes in depression levels. Key components of intervention modules based on social cognitive theory include self-efficacy, outcome expectations, social support, and goal setting.

2. LITERATURE REVIEW

2.1 Epidemiological Models

Epidemic is defined as a widespread occurrence of an infectious disease in a community at a particular time. The key research questions arise in modelling epidemics are the risk of an epidemic to occur, the severity level of the epidemic, duration, and what impact a particular intervention have on the risk, severity and duration of the epidemic.

The basic reproductive ratio R_0 is an important epidemiological measure for how infectious a disease is. It's defined as the average number of people an infectious person will infect, assuming that the rest of the population is susceptible. A threshold of 1 is used to determined whether disease will die out or can explode.

Epidemiological studies that divide a population into compartments are called compartmental models. The most commonly used epidemiological models are SIR and SEIR, which have been widely used in many studies analyzing the spread of infectious disease including Ebola and COVID-19 (Cho, 2016; Chowell and Nishiura, 2014; Kumar et al., 2021; Abou-Ismail, 2020). Epidemiological models utilize differential equations to focus on the rate of change of variables as time passes. This concept of compartmental modeling is based on a birth and death process, which is a continuous Markov process (Kermack, 1927; Ross, 2014). In Kermack (1927), it was explained that the fundamental theory of epidemic modeling involves the process of infection, transitioning to recovery or death among individuals, and how these events change over a specific unit of time. Additionally, Kermack (1927) explained the necessary parameters such as the infectivity, recovery, and death rates for epidemic modeling. This compartmental model has been applied to answer following questions: (i) how many individuals will be infected and died; (ii) How long will the epidemic last; and (iii) How much good would mitigation strategies do in reducing the severity of the epidemic (Kermack, 1927; Shin, 2021; Brauer et al., 2008). During last decades, the SIR or SEIR model has been applied to

model epidemic dynamics such as transmission of Ebola, Mers, influenza, and Severe Acute Respiratory Syndrome (SARS) (Brauer et al., 2008; Chowell and Nishiura, 2014; Chowell et al., 2003; Diaz et al., 2018; Lin et al., 2023; Beckley et al., 2013). Several researchers have enhanced the classical compartmental models by adding variables that control the rates of changes between compartments or by adding more compartments to specify diverse conditions of patients.

Since compartmental modeling has been considered as an application of the birth and death process in Markov process (Ross, 2014), understanding the fundamental assumptions of Markov process is useful in epidemic modeling. Definitions and assumptions of Markov process is described as follows.

2.1.1 Markov process

Let $\{X_n, n = 0, 1, 2, ..., \}$ be a stochastic process that takes on a finite countable number of possible values. If $X_n = \mathrm{i}$, then the process is said to be in state i at time n. Suppose that whenever the process is in state i, there is a fixed probability P_{ij} that will next be in state j. That is, suppose that

$$P\left(X_{n+1} = \mathrm{j} \mid X_n = \mathrm{i}, X_{n-1} = \mathrm{i}_{n-1}, ..., X_1 = \mathrm{i}_1, X_0 = \mathrm{i}_0\right) = P_{\mathrm{ij}} \tag{2.1}$$

for all states i_0 , i_1, ... , i_{n-1}, i, j and all $n \geq 0$. This stochastic process is defined as a Markov chain (Ross, 2014). An assumption defining a Markov chain is that the conditional distribution of X_{n+1} given all the past events X_0, X_1, ..., X_{n-1} depends on these past events only through the event at the end of day n. P_{ij} represents the probability that the process will, when in state i, next make a transition into state j. Therefore, the properties of P_{ij} include

$$P_{\mathrm{ij}} \geq 0, \qquad \mathrm{i}, \mathrm{j} \geq 0; \qquad \sum_{\mathrm{j}=0}^{\infty} P_{\mathrm{ij}} = 1, \qquad \mathrm{i} = 0, 1, ... \tag{2.2}$$

A continuous-time Markov chain is a stochastic process { X(t), t≥0 }, if for all $s, t \geq 0$ and integers must be non-negative $\mathrm{i}, \mathrm{j}, x(u)$, and $0 \leq u < s$ as defined in 2.3

$$\begin{aligned} P\{X(t+s) = \mathrm{j} | X(s) = \mathrm{i}, X(u) = x(u), 0 \leq u < s\} \\ = P\{X(t+s) = \mathrm{j} | X(s) = \mathrm{i}\} \end{aligned} \tag{2.3}$$

A birth and death process is a continuous-time Markov chain with states {0,1,...} for which transitions from state n may go only to either state $n-1$ or state $n+1$. This process describes the dynamics of population size in an epidemic. Epidemic modeling considers the change of population size from a state (e.g., infected), say n, to a connected state (e.g., exposed, recovered), say $n-1$ or $n+1$, at time t. Further, the rate of change in population size at time t only considers the population size at time $t-1$, which is similar with the assumption of Markov process. Particularly, the key epidemiological quantity, the basic reproduction number R_0, is used to identify situations when a disease can invade a population with an assumption that the dynamics of host births and deaths are not considered (Kermack and McKendrick, 1927). If each member acts independently of the others and takes an exponentially distributed amount of time, with mean latent time $1/\lambda$, to infect another member, then if $X(t)$ is the population size at time t, then $X(t), t \geq 0$ is a pure birth process with $\lambda_n = n\lambda,\ n \geq 0$ (Ross, 2014). It shows that the dynamics of the second infection is evaluated based on the birth and death process.

2.1.2 Compartmental modeling

The classical compartmental modeling, say the SIR or SEIR modeling, has been modified by introducing additional compartments or by dividing a compartment into multiple sub-compartments. The more compartments are added to the model, the more parameters are required to explain the rates of change from one compartment to another. For instance, Enns et al. (2020) divided the susceptible population by nine 10-year age

groups, while Nabi (2020) added 4 more compartments to the SEIR model; including a symptomatic infected-, an asymptomatic infected-, a quarantined-, and the hospitalized compartment. Similarly, Friji et al. (2021) added a quarantine- and hospitalization compartment to the SEIR model. Shin (2021) modeled the COVID-19 epidemic in Korea for the one-year period from February 18 2020 to February 8 2021. The study employed the SEIR and SEIRD models to estimate time-varying and context-dependent parameters of the epidemic along with multiple stages of its development. Findings in Shin (2021) showed that the government's effective non-pharmaceutical interventions significantly reduced transmission rate and the basic reproduction number. The study discussed that the compartment of D in the basic SIR model improved the analytical robustness.

Susceptible-Infectious-Removed (SIR) model

One of the simplest epidemiological models is the Susceptible-Infectious-Removed (SIR) model. The SIR model divides a population into three compartments – Susceptible-Infectious-Removed. It measures the number of people in each group changes. For instance, the number of susceptible population decreases as the epidemic spreads among people, whereas the number of infectious people increases. These changes can be modeled by a differential equations compartmental model.

Let S(t) denote the number of susceptible people at time t. Let I(t) denote the number of infectious individuals at time t. Similarly, R(t) denote the number of removed individuals at time t. Removed means that they are no longer infectious because they recovered or died. Let N denote the total number of people. Then,

$$S(t) + I(t) + R(t) = N$$

We can rewrite it as

$$\frac{S(t)}{N} + \frac{I(t)}{N} + \frac{R(t)}{N} = 1$$

and, let $s(t)$, $\mathrm{i}(t)$, and $r(t)$ denote each composite function, respectively.

The sum of the three variables at time t remains constant as long as the total number of people in the population N is constant. The SIR model assumes that a closed population of constant size N. If there is a significant change in N, the SIR model cannot be used (Abou-Ismail, 2020).

Intuitively, $S(t)$ will either stay in $S(t)$ or move into the $I(t)$. As a result, the rate at which susceptible individuals (S(t)) get infected must be negative. The magnitude of this change depends on the ratio of infected people at t, $\mathrm{i}(t)$, the ratio of susceptible people, $s(t)$, and the likelihood of disease transmission between the two groups, β. Then, the rate of change of the susceptible individuals over time can be expressed as:

$$\frac{dS}{dt} = -\frac{\beta S(t) I(t)}{N}$$

When β is relatively large, the infection spreads fast and $S(t)$ decreases quickly, whereas when β is relatively small, the disease spread becomes slower.

Let γ denote the rate of change over time of $I(t)$. It can be interpreted as the rate at which $I(t)$ moves into $R(t)$. Then. the rate of change of the $I(t)$ depends on both γ and β. It also depends on the ratio of individuals in the infectious and susceptible groups at time t. It can be mathematically expresses as

$$\frac{dI}{dt} = \beta S(t) I(t)/N - \gamma I(t)$$

As β increases, the $I(t)$ increases, and as γ increases, the $I(t)$ decreases.

The rate of change of $R(t)$ at time t depends on $I(t)$ and γ.

$$\frac{dR}{dt} = \gamma I(t)$$

The average number of days it takes for an individual to recover from the disease, denoted by n, is inversely proportional to γ. Note that the rate of change of the removed group is always positive, since $R(t)$ can only increase with time. When γ is large, people recover very quickly and more from the $I(t)$ to the $R(t)$. This means that the disease can be under control.

The basic reproduction number R_0 is defined as the average number of secondary cases generated by a primary case over his/her infectious period when introduced into a large population of susceptible individuals (Diekmann et al., 1990). R_0 is an estimate of the epidemic growth at the start of an outbreak if everyone is susceptible (Cho, 2016; Chowell et al., 2004). That is, R_0 is used to describe the contagiousness or transmissibility of infectious agents. Assuming that the total population is 1.0 and that each of the three subgroups are a fraction of the total, R_0 can be calculated as:

$$R_0 = \frac{\beta}{\gamma}$$

Susceptible-Exposed-Infectious-Removed (SEIR) model

The Susceptible-Exposed-Infectious-Removed (SEIR) model is one of the variants of the SIR model. It's distinctive characteristic is that the SEIR model considers the exposed group, which has a latent period. In other words, the exposed group $E(t)$ is a group between the $S(t)$ and the $I(t)$. It includes individuals who have been exposed to the infection but are not infectious yet. This additional group enables the model to be more realistic when simulating the infectious disease because most of infectious disease

have latent period to develop the symptoms. For instance, Ebola has about 2 to 21 days of latent period and COVID-19 has about 5.6 days of latent period after contact.

The rate of change of $S(t)$ and $R(t)$ remains same in the SEIR model as the SIR model. We introduce the likelihood that an exposed person becomes infected, δ. Then, the rate of change of $E(t)$ is mathematically expressed as:

$$\frac{dE}{dt} = \beta S(t)I(t)/N - \delta E(t)$$

and, accordingly the rate of change of $I(t)$ becomes:

$$\frac{dI}{dt} = \delta E(t) - \gamma I(t)$$

The R_0 measures the initial growth rate of the epidemic and for the model above it can be shown that $R_0 = \beta_0/\gamma$, where β_0 is the pre-interventions transmission rate and $1/\gamma$ is the mean infectious period. The effective reproductive number at time t, $R_{eff}(t) = (\beta(t)/\gamma)s(t)$, measures the average number of secondary cases per infectious case t time units after the introduction of the initial infections and $s(t) = \frac{S(t)}{N} \approx 1$ as the population size is much larger than the resulting size of the outbreak. Therefore, $R_{eff}(0) = R_0$ (Chowell et al., 2004).

The classical SIR model was used to estimate the dynamics of generation of misinformation (Cinelli et al., 2020). Multiple social media platforms were used to compare misinformation pattern across different sources. Cinelli et al. (2020) described the SIR model by a set of differential equations:

$$\begin{aligned}
\frac{dS}{dt} &= -\beta S(t)I(t)/N \\
\frac{dI}{dt} &= \beta S(t)I(t)/N - \gamma I(t) \\
\frac{dR}{dt} &= \gamma I(t)
\end{aligned}$$

where $S(t)$ is the number of susceptible, $I(t)$ is the number of infected and $R(t)$ is the number of recovered. The study interpreted the number $I(t) + R(t)$ as the number of authors that have published a post on the subject. Least square estimation was used to estimate the models' parameters and bootstrapping was used to get the range of parameters. As a result, each platform has $R_0 > 1$, which implies the possibility of an infodemic. Therefore, when making intervention strategy using social media in a crisis, it is important to choose a social media platforms by considering their contexts (Cinelli et al., 2020).

Kumar et al. (2021) studied social media effects in reducing transmission rate of influenza and pandemic by using SEIR model. The researchers used decreasing functions with respect to the current number of infected individuals in the population to incorporate the number of tweets related to COVID-19. Specifically, the study used the total number of tweets (M(t)) about the infectious disease at any given time. In the SEIR model, the researchers divided the susceptible individuals into two groups; who are not influenced by the tweets (S) and who are influenced by the tweets (S_1).

Individuals who are influenced by the tweets at time t will move to S_1 at the rate of $\tau M(t)$. The transmission rates β and β_1 are the rates at which a susceptible individual in S and S_1 is infected by infectious individuals, respectively. N is the total population. The SEIR model is described by following equations:

$$
\begin{aligned}
\frac{dS}{dt} &= -\frac{\beta I(t)}{N} S(t) - \tau M(t) S(t) \\
\frac{dS_1}{dt} &= -\frac{\beta_1 I(t)}{N} S_1(t) + \tau M(t) S(t) \\
\frac{dE}{dt} &= \frac{\beta I(t)}{N} S(t)) + \frac{\beta_1 I(t)}{N} S_1(t) - \sigma E(t) \\
\frac{dI}{dt} &= \sigma E(t) - \gamma I(t) - \delta D \\
\frac{dR}{dt} &= \gamma I(t) \\
\frac{dD}{dt} &= \delta D(t)
\end{aligned}
$$

2.1.3 Parameter estimation

Once the transmission of infectious diseases is represented using compartmental models, it becomes essential to ascertain the parameter values that indicate the rates at which individuals transition between compartments. These parameter values are crucial for examining and predicting disease dynamics, particularly in the context of disaster management. Typically, these parameter values are determined by comparing the reported number of infected cases by organizations such as WHO or CDC with the predicted number of infected cases from a proposed model. A good model is expected to show a small discrepancy between the reported cases and the analytical results. Hence, one of the primary tasks in epidemic modeling is to identify the optimal or best parameters for these explanatory models, ensuring they accurately capture the real-world phenomenon. Chowell et al. (2003) studied SARS outbreaks using a compartmental model. In the process of data fitting through simulation, the research varied two parameters: a relative measure of reduced risk among diagnosed cases and a rate of progression from infected to diagnosed per day. Meanwhile, all other parameters were either roughly estimated based on collected data or literature, or fixed arbitrarily. Yarsky (2021) es-

timated parameters of an SEIR model for coronavirus using a genetic algorithm with a multi-objective function consisting of residuals in both infected case data and casualty counts in fitting. Nabi (2020) utilized a trust-region-reflective algorithm to modify the baseline parameters and discovered the optimal parameters for their compartmental model containing 8 compartments. Tsay et al. (2020) used the least-squares regression for parameter estimation to find the parameter values that minimize the mean squared error (MSE) between the predicted cases and measured values of the total infected-, recovered-, and dead subjects reported by Johns Hopkins University. The study used the Pyomo package in Python to find the optimal parameter values with the minimum MSE. Then, the study used an interior-point filter line-search (IPOPT) algorithm to solve the nonlinear dynamic optimization problem (Wächter and Biegler, 2006). Nsoesie et al. (2013) utilized the simulation optimization approach for forecasting the influenza epidemic curve. For parameter estimation, Nsoesie et al. (2013) applied the Nelder-Mead simplex method to find the optimal parameter set that minimizes the error between the estimated infected counts and the true infected counts. He et al. (2007) employed both the particle swarm optimization method and the genetic algorithm to estimate parameters for chaotic systems, including disease transmission. In He et al. (2007), the particle swarm optimization method revealed better accuracy in parameter estimation in their study.

One important assumption made in Kumar et al. (2021) was the vital intervention in controlling COVID-19 was social distancing and the study borrowed the estimation of its effect from Chu et al. (2020). The social distancing measures could decrease the transmission risk by 7.5 - 15.9% (i.e. $\eta \in (7.5\%, 15.9\%)$) and the study assumed that $\beta_1 = \beta(1-\eta)$, where β is the transmission rate at which a susceptible individuals in non-social media users. Then, this study solved an optimization problem:

$$min|\hat{SV} - 2.1\hat{S_1V}|$$

subject to

$$16.1\%N \leq \hat{S}_1 \leq 21\%N$$
$$14.3\%N \leq \hat{V} \leq 14.4\%N$$
$$v_1 \geq v \geq 0$$

Then, the study searched the solution space by running the

$$\frac{dS}{dt} = -\frac{\beta I(t)}{N} S(t) - \tau M(t) S(t) - vS$$

where v denotes vaccine rate, which are both set to 0. Then, the solution sets of when $\eta = 16\%, 21\%$ are combined and are simulated as long as the objective value is less than 0.5 (Kumar et al., 2021). Kumar et al. (2021) assumed the total population N is 10,010 and the susceptible population S(0) is 10,000. Next, ninety four keywords and hashtags are used to collect tweets, such as corona, coronavirus, covid, covid19, covid-19, sarscov2, sars cov2, quarantine, flatten the curve on April 18, 2020 and May 16, 2020. The authors then adjusted the raw data to make the daily number of tweets consistent during the time period from March 22, 2020 to July 20, 2020. After normalizing, the daily normalized number of tweets is used as M(t). Kumar et al. (2021) use the performance measures to evaluate the effectiveness of social media in the COVID-19 pandemic: (1) peak time when the infected is at its maximum, (2) peak magnitude, which is the number of people who are infected at the peak time, (3) total infected, and (4) the total deaths caused by COVID-19.

In the findings, the impact of social media seems to have the least influence on peak times. Social media's effects on the total number of infections and deaths are more responsive to changes in the basic reproduction number (R_0) compared to its impact on the magnitude of the peak. Specifically, when R_0 exceeds 1.9, the declines in

total infections and deaths attributed to social media diminish more rapidly than the reductions in peak magnitude. In general, social media proves less effective when dealing with either mild or extremely severe infectious diseases. However, it is most effective in curtailing a pandemic when the disease's R_0 falls within the range of 1.5 to 1.9 (Kumar et al., 2021). Moreover, Kumar et al. (2021) found that the peak time can be prolonged or shortened when social media is in effect. Peak magnitude, total infected cases, and total deaths can be reduced when social media is in effect. In conclusion, social media is less effective when the infectious disease is mild or very severe, and social media is most effective in mitigating the pandemic if the disease's R_0 is between 1.5 and 1.9. Thus, Kumar et al. (2021) found that social media has a positive effect in mitigating the infectious disease.

2.2 Artificial Intelligence

The definition of Artificial Intelligence (AI) was defined as the science and engineering of making intelligent machines (McCarthy et al., 1955). In fact, AI has been implemented in many real-world applications including production systems and healthcare systems (Christopher Manning, 2020; Ivanov et al., 2021; Yang et al., 2019; Pham et al., 2020). The concept of AI gained attention after a five-game match of the game Baduk (or Go) between an AI program developed by Google DeepMind, called AlphaGo, and Saedol Lee, the world champion, in March 2016 (Korea Baduk Association, 2016; BBC, 2016). AlphaGo defeated Saedol Lee with 4 to 1 score. Through this DeepMind Challenge Match, Google demonstrated the complicated strategies and computational power of its AI machine. AlphaGo used Monte Carlo tree search ensembled with neural networks. The concept of neural networks was inspired by biological systems, particularly by how neurons in the brain might work (McCulloch and Pitts, 1943; Widrow et al., 1960; Rosenblatt et al., 1962; Rumelhart et al., 1986). AlphaGo was able to consider 250^{150} or 10^{360} moves due to its neural network architecture and due to the advancement

of hardware such as processors, memory, and storage (Sze et al., 2017). Within the realm of AI, the domain of machine learning is regarded as an analytical technique and methodology. Therefore, machine learning is considered as an application of AI (Microsoft Azure, 2024). This research includes the development and application of machine learning methodologies in prediction and recommendation. Specifically, two foundational techniques in machine learning, support vector machine and K nearest neighbor, are discussed.

2.2.1 Machine learning

The problem of searching for patterns in data is a fundamental research area (Bishop and Nasrabadi, 2006). Efforts to understand data requires that researchers classify and categorize different data sets to understand data and generalize that understanding to new data (Bishop and Nasrabadi, 2006). To achieve this goal, machine learning is applied to given data to learn (train) and to predict (test). The prime theory behind machine learning is based on Bayes' theorem (Equation 2.4), which deduces the likelihood of a new event (Y) based on existing knowledge or empirical evidence (X).

$$\begin{aligned} p(Y|X) &= \frac{p(X|Y)p(Y)}{p(X)}, \\ p(X) &= \sum_{Y} p(X|Y). \end{aligned} \tag{2.4}$$

Three types of problems exist in machine learning: supervised, unsupervised, and reinforcement learning. Supervised learning is a problem where the goal is to take an input vector x and to assign it to one of K discrete classes C_k where $k = 1, ..., K$. Each class is disjoint, so that each input is assigned to one and only one class. Examples of classification methods are support vector machines (SVM) and k-nearest neighbor (KNN). On the other hand, the problem of clustering, or unsupervised learning, is to

discover groups of similar examples within the data. Lastly, the goal of reinforcement learning is to find actions to maximize a reward. Considering the research scope in this dissertation, methods in solving supervised learning to be applied to semi-supervised learning are discussed in the following sections.

2.2.2 Support vector machine

Support vector machines (SVM) are useful for classification problems when two data sets are not linearly separable. For instance, two sets of points in $\mathbb{R}^n$ are given by $\{x_1, ..., x_N\}$ and $\{y_1, ..., y_M\}$. The goal is to find a function $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$\begin{aligned} f(x_\mathrm{i}) &> 0, \quad \mathrm{i} = 1, ..., N, \\ f(y_\mathrm{i}) &< 0, \quad \mathrm{i} = 1, ..., M, \end{aligned} \tag{2.5}$$

and f classifies or separates the two sets of points. In linearly separable problems, the goal is to find an affine function, or hyperplane, $f(x) = a^T x - b$ such that

$$\begin{aligned} a^T x_\mathrm{i} - b &> 0, \quad \mathrm{i} = 1, ..., N, \\ a^T y_\mathrm{i} - b &< 0, \quad \mathrm{i} = 1, ..., M. \end{aligned} \tag{2.6}$$

However, when the two sets of data cannot be linearly separable, support vector machines are useful for approximate linear classification (Boyd and Vandenberghe, 2004). Support vector machines add nonnegative slack variables to a feasibility problem to increase the margin around the boundary, or the supporting hyperplane, between two sets of data.

$$\begin{aligned} a^T x_\mathrm{i} - b &> 1 - u_\mathrm{i}, \quad \mathrm{i} = 1, ..., N, \\ a^T y_\mathrm{i} - b &< -(1 - v_\mathrm{i}), \quad \mathrm{i} = 1, ..., M. \end{aligned} \tag{2.7}$$

The goal of support vector machines is to find a, b, and nonnegative u and v that satisfy the inequalities (Equation 2.7). To solve this, a heuristic (Equation 2.8) can be minimized.

$$\begin{aligned} \min \quad & 1^T u + 1^T v \\ \text{subject to} \quad & a^T x_{\mathrm{i}} - b > 1 - u_{\mathrm{i}}, \quad \mathrm{i} = 1, ..., N, \\ & a^T y_{\mathrm{i}} - b < -(1 - v_{\mathrm{i}}), \quad \mathrm{i} = 1, ..., M \\ & u \geq 0, \quad v \geq 0. \end{aligned} \tag{2.8}$$

Support vector machines are used in this research to be selected as a classifier in self-training machine learning to predict the most likely preventive behavior with which each tweet user complies. Specifically, the support vector machines with the linear kernel is used. In text data, they are unstructured with many features and entries. The dimension of textual data should be reduced instead of mapping data into high dimensional space, which is not useful for already high dimensional data (Hsu et al., 2003) due to expensive computation cost and low accuracy. Support vector machines with linear kernel allow fast computation as well as to overcome over-fitting. The objective function for support vector machines with linear kernel is mathematically formulated as

$$\min_{w,b} \frac{1}{2} w^T w + C \sum_{\mathrm{i}=1}^{n} max(0, 1 - y_{\mathrm{i}}(w^T \phi(x_{\mathrm{i}}) + b)), \tag{2.9}$$

where w, b are the model parameters of the hyperplane, x represents the input variables, C is a constant, and ϕ is the identity function (i.e., $f(x) = x$). All x's are vectorized to real numbers. The distance from the hyperplane to input data shows how confident prediction is. Only if an x is classified correctly and the distance from the plane is larger than the margin will there be no penalty. The purpose of penalty function is to assure the classification is correctly made, minimizing misclassification. The Euclidean

distance is proper to measure the closeness (or absolute differences) between text and the decision boundary (Bertsimas and Tsitsiklis, 1997).

2.2.3 K-nearest neighbor

K-nearest neighbor algorithm (KNN) classifies the new data based on the proximity between the new data and existing training data (Dasarathy, 1991). K-nearest neighbor algorithm is useful particularly for multiple recommendation or multi-label classification. The K-nearest neighbor algorithm finds the k closest data points based on a distance measure between the new data and the training data. The new data is classified to the most frequent class of its k nearest data (Latah and Toker, 2020). Possible distance metrics include l_1 norm (i.e., Manhattan distance), l_2 norm (i.e., Euclidean distance), l_∞ (i.e., Chebyshev distance), etc. The formula of each distance metric is expressed in Equation 2.10.

$$\sqrt[\lambda]{\sum_{\mathrm{i}=1}^{n} |x_\mathrm{i} - y_\mathrm{i}|^\lambda}, \tag{2.10}$$

here x_i is a point of the vector x, whereas y_i is a point of the vector y.
The performance of k-nearest neighbor algorithm is affected by the distance metric with the choice of parameter k. When $\lambda = 1$, the metric is the Manhattan distance. If $\lambda = 2$, the metric is the Euclidean distance. If $\lambda = \infty$, the metric is the Chebyshev distance. Selection of a distance metric depends on the performance measures used in research of interest.

In applying k-nearest neighbor algorithm to the research question in this study, cosine distance is selected and the optimal k is selected by comparing the F-1 scores. Cosine similarity is a popular metric to measure the similarity between texts. As the cosine distance considers the angle between two texts instead of length, it describes the context similarity.

K-nearest neighbor algorithm was applied to twitter data in this study. Suppose that an input tweet includes a similar set of words in the body text to other tweets that have been labeled. The input tweet can be considered as similar to the labeled tweets. This is a classification problem. Specifically, when similar contextual information is present, characteristics (e.g., prevention behaviors) in the input data can be inferred from the labeled tweets. Similarity between an input tweet (x) and another tweet (x') can be measured by calculating the difference of angle between two vectors (i.e., cosine distance). The distance, or dissimilarity, is opposite to the similarity. These two measures are mathematically expressed below.

$$\begin{aligned} sim(x, x') &= cos(\theta) = \frac{x \cdot x'}{\|x\| \cdot \|x'\|}, \\ dist(x, x') &= 1 - sim(x, x') \end{aligned} \tag{2.11}$$

where θ is an angle between two vectors $\vec{x}, \vec{x'}$. The problem is to find a set of the n nearest neighbors of x as $X' = \{x'_1, x'_2, ..., x'_n\} \subseteq D$ such that

$$\begin{aligned} &\forall (x', y') \in D \setminus X', \\ &\qquad dist(x, x') \geq \max_{(x'', y'') \in X'} dist(x, x''), \end{aligned} \tag{2.12}$$

that is, every point in D but not in the set X' is at least as distant from x as the furthest point in X'. Then, K-nearest neighbor model chooses the most common label, say y, in X' (Cover and Hart, 1967).

2.2.4 Semi-supervised learning

Semi-supervised learning is a methodology that combines the techniques of supervised learning and unsupervised learning. When a training set does not have enough data, the learnable features are insufficient to allow for effective learning processes. And

the process of labeling data is expensive, requiring expert knowledge. Semi-supervised learning alleviates the need for labeled data by allowing a model to leverage unlabeled data (Foulds and Smyth, 2011; Berthelot et al., 2019).

One fundamental methodology of semi-supervised learning method is generative mixture models. This assumes a generative model $p(x, y) = p(y)p(x|y)$ where $p(x|y)$ is an identifiable mixture distribution, for example a Gaussian mixture model. With a large amount of unlabeled data, the mixture components can be identified, then theoretically, only one labeled example per component is needed to fully determine the mixture distribution (Zhu, 2005). Another fundamental theory for semi-supervised learning is self-training. In self-training, a classifier is first retrained with a small amount of labeled data. The classifier is then used to classify the unlabeled data. Typically the most confident unlabeled points, together with their predicted labels, are added to the training set. The classifier is re-trained and the procedure repeated (Rahmani and Goldman, 2006; Zhu, 2005).

2.3 Text Analytics

Text analytics, a technique of natural language processing, is tools, techniques, and algorithms to process and understand natural language-based data (i.e., text), which is unstructured. Text analytics relies heavily on machine-readable dictionaries. Natural language toolkit (NLTK) is the most popular dictionary for machines (Bird et al., 2009) and WordNet is one of the NLTK corpus readers. WordNet provides data entries that have traditional lexicographic information and a programming-familiar structure (Miller, 1995). Several text analytics packages in python depend on WordNet lexical databases, especially to English texts.

Linguistic relations were defined in WordNet so that machines can identify relations among words. For instance, ssemantic relations includes synonymy (similar), antonymy (opposite), hyponymy (subordinate), meronymy (part), troponomy (manner), and en-

tailment. A syntactic category includes English nouns, verbs, adjectives, and adverbs and past study has demonstrated that adjectives are good indicators of subjective, evaluative sentences (Miller, 1995; Esuli and Sebastiani, 2006). More than 116,000 of these semantic relations between words and word senses were included in the original WordNet. Additionally, WordNet considers contextual representations for machine translation. It provides alternative senses of a word so that a computer distinguishes between different sets of linguistic contexts. Several methods in text analytics have been investigated before determining the appropriate method to address the research question.

2.3.1 Text classification

Text classification is defined as the process of assigning text documents into one or more classes or categories, given a predefined set of classes (Sarkar, 2016; Mirończuk and Protasiewicz, 2018). From perspective of machine learning, text classification is a supervised learning. In the text classification process, the following analysis steps are involved - data acquisition, data preprocessing including labeling and normalization, feature selection, training, and evaluation (Mirończuk and Protasiewicz, 2018). A conceptual representation of the text classification process is shown in Figure 2.1 (Sarkar, 2016).

2.3.2 Topic modeling

Topic modeling categorizes given documents (in this case groups of texts) into similar topics based on the words that constitute the documents. Identifying relevant topics in large conversational data such as social media data is often difficult (Baeza-Yates, 1999). However, analyzing a large set of documents is time consuming, and requires efficient dimensionality reduction techniques (Deerwester et al., 1990).

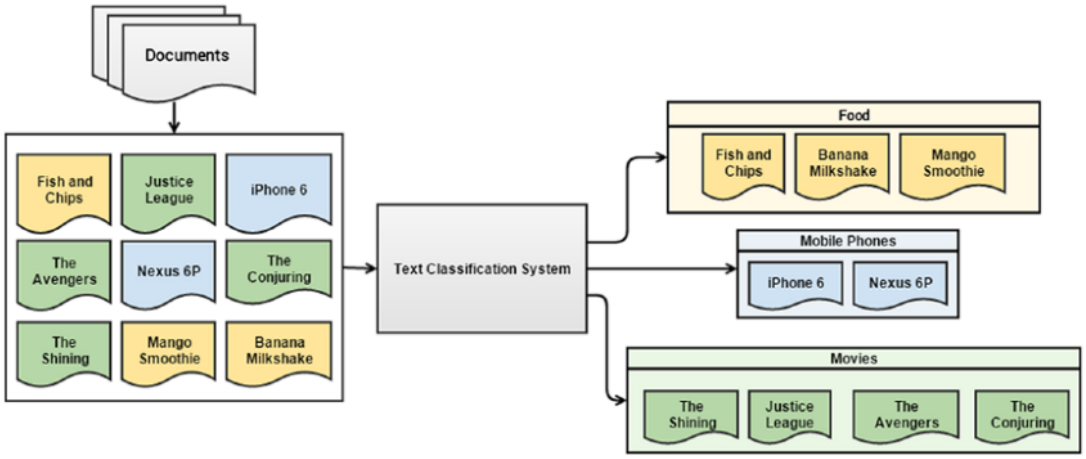


Figure 2.1. Conceptual overview of text classification

One of the fundamental dimensionality reduction techniques, or topic modeling methods, is Latent Dirichlet Allocation (LDA). The objective of the LDA model is to find the best representative document-topic distribution and topic-word distribution. To do so, the LDA model evaluates 2 probabilities for every topic: (i) proportion of words in the document(D) that are currently assigned to the topic(t); and (ii) the proportion of documents, in which the word is also assigned to the topic(t).

LDA is a generative, unsupervised, and probabilistic topic modeling method used for discovering and extracting the hidden structure topics in textual data (Blei et al., 2003; Ghosh and Guha, 2013; Griffiths and Steyvers, 2004; Murshed et al., 2022). Generative models can be used to predict complex latent structures related to a set of language-based observations, enabling to use statistical inference to recover this structure (Griffiths and Steyvers, 2004). For instance, text includes the observed data (i.e., words) which are intended to communicate a latent structure (i.e., implicit meaning) consisting of a set of topics (Blei et al., 2003; Griffiths and Steyvers, 2004).

The LDA model decomposes the observed documents of words (i.e., the larger matrix) into two sub-matrices: the document-topic matrix and the topic-word matrix. LDA seeks to find a probability distribution of a mixture of topics as well as a probability distribution of a set of words for each topic. Figure 2.2 visualizes the decomposition process of LDA, where D represents a document, w represents a word, and T represents a topic. The number 1 means included, whereas 0 means not included.

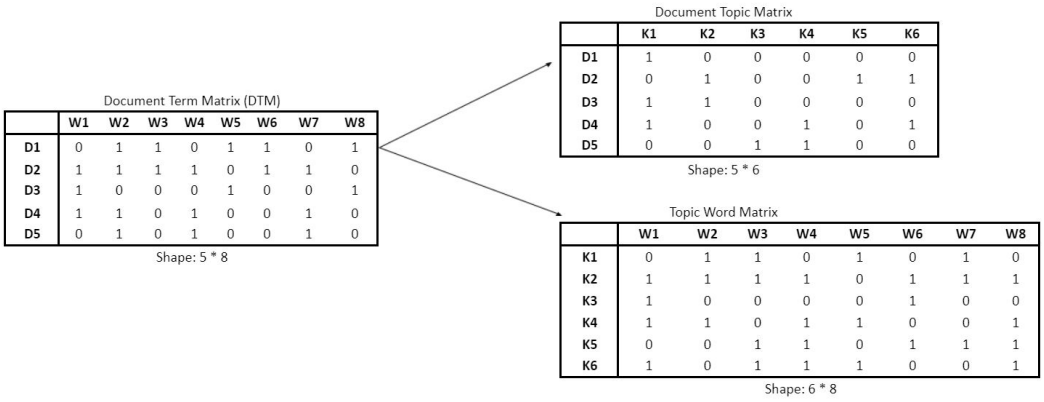


Figure 2.2. LDA and decomposition of document-word matrix into document-topic matrix and topic-word matrix (Seth, 2021)

The generative process of the LDA method is iterative as described in Algorithm 2.3. If there exist T topics, the probability of the ith word in a given document is

$$P(w_{\mathrm{i}}) = \sum_{\mathrm{j}=1}^{T} P(w_{\mathrm{i}}|z_{\mathrm{i}} = \mathrm{j})P(z_{\mathrm{i}} = \mathrm{j}), \tag{2.13}$$

where z_{i} is a latent variable indicating the topic from which the ith word was drawn and $P(w_{\mathrm{i}}|z_{\mathrm{i}} = \mathrm{j})$ is the probability of the word w_{i} under the jth topic. $P(z_{\mathrm{i}} = \mathrm{j})$ gives the probability of choosing a word from topics j in the current document, which will vary across different documents.

LDA combines Equation 2.13 with a prior probability distribution on the probability of topics with a set of D multinomial distribution θ over the T topics. In LDA, documents

1: **procedure** LDA TRAINING
2: x ←texts ▷ Load data
3: x ←`preprocessing`(x) ▷ Tokenize, remove stop words, lemmatize
4: x ←`vectorizer`(x) ▷ Numerical representation of x
5: gensim.LDAModel(k)(): ▷ Training the LDA model
6: Choose θ_{i} ∼Dirichlet(α), i ∈ $\{1, ..., M\}$
7: Choose ϕ_k ∼Dirichlet(β), $k \in \{1, ..., K\}$
8: **for** i,j, where i ∈ $\{1, ..., M\}$, and j ∈ $\{1, .., N_{\mathrm{i}}\}$ **do**:
9: Choose a topic and word $t_{\mathrm{i,j}}$ ∼Multinomial(θ_{i}) & $w_{\mathrm{i,j}}$ ∼Multinomial($\phi_{t_{\mathrm{i,j}}}$)
10: LDAModel.print_topics() ▷ Display topics

Figure 2.3. Algorithm - Latent Dirichlet Allocation (LDA)

are generated by first picking a distribution θ over topics using a Dirichlet distribution, which determines $P(z)$ for words in that document. The words in the document are then generated by picking a topic j from this distribution and then picking a word from that topic according to $P(w|z = \mathrm{j})$, which is determined by a fixed $\phi^{(\mathrm{j})}$. The estimation problem becomes one of maximizing $P(w|\phi, \alpha) = \int P(w|\phi, \theta)P(\theta|\alpha)d\theta$, where $P(\theta)$ is a Dirichlet(α) distribution. The integral in this expression is intractable, and ϕ is thus usually estimated by using sophisticated approximations.

The Dirichlet distribution is Bayesian-alike, because its conjugate prior is also a multinomial distribution (Ferguson, 1973). The definition of conjugate prior is if the posterior distribution $p(\theta|x)$ is in the same probability distribution family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function $p(x|\theta)$ (Raiffa et al., 1961). The LDA model uses the Dirichlet distribution in an iterative process to identify similarity, frequency, and co-occurrence of topics and words for the given documents.

2.3.3 Sarcasm detection

Indirect speech is a sophisticated form of a speech act in which speakers convey their message in an implicit way (Tsur et al., 2010). Sarcasm is a manipulation of indirect speech and communicates the opposite message of what the speaker is saying (Bharti et al., 2015). Therefore, contextual information could be interpreted incorrectly if speech is not examined for sarcastic meaning in context n analysis.

Given the political discourse that surrounded COVID-19, sarcasm in tweets must be considered. Tsur et al. (2010) studied sarcasm in user reviews on Amazon.com using semi-supervised pattern acquisition and a classification algorithm. In their algorithm, syntactic and pattern-based features were employed. Syntactic features were directly related to linguistics such as grammatical features. Pattern-based features were composed of high frequency words appearing more than 1,000 words per million and content words less than 100 words per million (Tsur et al., 2010; Davidov and Rappoport, 2006) (Figure 2.4).

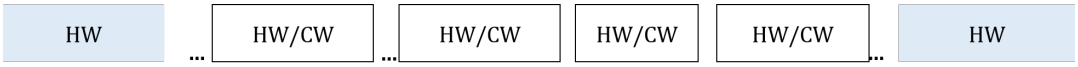


Figure 2.4. Pattern extraction

Tsur et al. (2010) defined a decision rule for pattern extraction: high frequency words (HW) were placed in the first and the last position in the pattern. Also, 2 to 6 positions were assigned to high frequency words and 1 to 6 positions were assigned to content words (CW). The study included punctuation characters in high frequency words.

Extracted patterns were filtered by removing product-specific patterns and patterns clearly showing both sarcastic and not sarcastic speech (Tsur et al., 2010). Speech patterns were used to match each sentence. That is, if all the pattern components appeared in the sentence without additional words, then the sentence was assigned

by the value 1. If some additional non-matching words were inserted in the pattern components, α was assigned to the sentence. If some number of N pattern components appeared in the sentence, and some non-matching words were inserted, $\gamma \times \frac{n}{N}$ was assigned. If nothing or only a single pattern component appeared in the sentence, then 0 was assigned. The study used $\alpha = \gamma = 0.1$ ($0 \leq \alpha \leq 1$ and $0 \leq \gamma \leq 1$ can be used). The study used k-nearest neighbors (KNN)-like methods with Euclidean distance to identify sarcasm (Tsur et al., 2010) with an average value of weighted labels based on frequency in classification.

Bamman and Smith (2015) modeled the relationship between a tweet and an author's past tweets in order to improve accuracy of sarcasm detection. The study collected 3,200 tweets from authors who mentioned #sarcasm or #sarcastic in the Gardenhose sample of tweets from August 2013 to July 2014. The study sub-sampled this set to include only tweets that were in response to another tweet. This yielded a positive training set of 9,767 tweets. The reason why the number of subsamples was greater than the collected tweets was not specified. The author selected an equal number of tweets from users over the same time period who had not mentioned #sarcasm or #sarcastic in their messages. Bamman and Smith (2015) used binary logistic regression with ℓ_2 regularization using 10-fold cross-validation, split on authors. Five combinations of four features were considered: tweet features only; tweet features and response features; tweet features and audience features; tweets features and author features; and all features together. For sarcasm detection, the author feature yield the greatest improvements in accuracy over the tweet features alone, whereas all feature classes display statistically significant improvements over the tweet only features. Additionally, the study found that the strongest audience based features do not show closeness between the author and audience.

2.3.4 Bidirectional encoder representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT) can also be used to model to extract sentiment. BERT is based on transformers, a machine learning technique for natural language processing (Devlin et al., 2018). Transformers are a type of neural network architecture, consisting of multiple layers. However, transformers do not have to process textual data with sequences in linear fashion. BERT alleviates the uni-directionality constraint (e.g., left to right or right to left language models) by a masked language model pre-training. Additionally, BERT was pre-trained by next sentence prediction to understand the relationship between two sentences. BERT was trained on the BooksCorpus and English Wikipedia (Seo et al., 2016). In fine-tuning, the self-attention in the transformer: (i) determines appropriate inputs and outputs; and (ii) encodes common text pairs (Seo et al., 2016; Devlin et al., 2018). BERT can be applied in a specific domain with an additional layer and fine-tuning. As BERT has been pre-trained on large corpus, this transfer learning scheme lowers cost when BERT needs to be find-tuned for domain-specific tasks.

2.3.5 Sentiment analysis

Sentiment analysis focuses on semantic inferences and enables researchers to identify the context (subjectivity) beyond the content (objectivity). Attributes of contents include more or less title, author, date and time, text body, location (may not be available), comments, interactions (e.g, likes, views, number of shared), etc. Characteristics of context is normally latent and not directly observable from the acquired text. Sentiment analysis actually focus on emotion recognition (Pang et al., 2008). Existing approaches to sentiment analysis can be grouped into three main categories: knowledge-based techniques, statistical methods, and hybrid approaches (Mirończuk and Protasiewicz, 2018; Cambria et al., 2013). Challenges of sentiment analysis lies in named-entity recognition,

ambiguity of words, and privacy. It is, therefore, easy to observe researchers grouped emotions into simple nuances - positive, neutral, and negative, rather than highly categorized emotions.

Zhou et al. (2021) suggested a framework to predict users' adoption behavior within different periods. The study extracted two features, users and contagions features based on a Latent Dirichlet Allocation (LDA) topic model and deep learning within a weighted network model for information diffusion. Attributes of contagions, including category, popularity, freshness, semantics, and sentiment affects the information spreading among connected users, while attributes of users, including social roles, preferences, and instant states, may affect their information adoption behaviors. Moreover, a sentiment-LDA topic model is used to represent both the semantic and sentiment features. A keywords vector is built according to the word frequency in descending order and selected the top 20% portion of keywords to reduce the computational burden. By the sentiment-LDA topic model, the words, the topics, and the sentiment polarities are generated. The author categorized the sentiment into three groups: negative, neutral, and positive. The authors then estimated the semantic topic and the sentiment polarity by the topic probability distributions. Finally, the authors generated the word during the sampling process with the semantic topic and the sentiment polarity. Therefore, the extracted topic set and the corresponding sentiment polarity set are utilized to represent the semantic and sentiment features in corpus contents (Zhou et al., 2021).

Subjective words in text documents are explored by unsupervised sentiment analysis. TextBlob is one of the python packages that is easily utilized in text analytics. TextBlob provides techniques to perform tasks such as part-of-speech tagging, noun phrase extraction, classification, translation, sentiment analysis, etc. The decision rules of sentiment analysis used in TextBlob are rule-based, meaning TextBlob refers to the predefined dictionary (e.g., the NLTK corpus), which contains pre-trained polarity score.

The outcomes of the TextBlob sentiment analysis include a polarity score, ranging from -1 to +1, and a subjectivity score, ranging from 0 to +1. The polarity score identifies if the tone of texts is negative or positive, whereas the subjectivity score distinguishes between factual information and subjective opinions.

The scoring mechanism is that any given text is broken down to words, called tokens, which are evaluated with the pre-trained scores (Steven, Loria , 2017; Miller, 1995). In determining the negative or positive sentiment, several linguistic relations and rules are used. For example, negation of a positive word flips the polarity score to negative by multiplying -0.5 to the original polarity score. Emphasis of a word weights the polarity score by $\times 1.3$.

The sentiment analyzer in TextBlob includes two sentiment analysis implementations, the pattern analyzer and the Naive Bayes analyzer (Loria, 2017). The pattern analyzer refers to a dictionary of adjectives and their labeled scores in the pattern library when evaluating polarity and subjectivity score of words in a given text. The pattern library takes the individual word scores from the sentiwordnet, a lexical resource (Esuli and Sebastiani, 2006).

On the other hand, the Naive Bayes analyzer is an NLTK classifier trained on a movie reviews corpus. The Naive Bayes analyzer is based on the Bayes rule to find the probability for a label, $P(\text{label}|\text{features})$. With the predefined dictionary with some labeled data, $P(\text{label})$ and $P(\text{features}|\text{label})$ are able to be found. The assumption is that all features are independent, given a label. That is,

$$P(\text{label}|\text{features}) = \frac{\begin{array}{c} P(\text{label}) \times P(\text{feature}_1|\text{label}) \times P(\text{feature}_2|\text{label}) \\ \times P(\text{feature}_3|\text{label}) \times \cdots \times P(\text{feature}_n|\text{label}) \end{array}}{P(\text{features})}$$

To build the conditional probabilities of $P(\text{feature}_\text{i}|\text{label})$, the Naive Bayes analyzer first chooses relevant words by removing the words with a probability less than a threshold. Then, for each word in the dictionary, the analyzer evaluates a probability of that word

being in each label. Finally, the classifier can make a prediction using the conditional probability. The probability of a label given features obtained from the Bayes rule is then used as a polarity score. In a sentence, these probabilities given multiple words are averaged, consequent on a single polarity score.

3. DATA

3.1 Scoping

The scope of study is limited to the infected area of interest, the United States of America, particularly New York. In the initial peak time of COVID-19 transmission, March 1 to May 31, 2020, New York had the most infected cases in the United States. Thus, the research scope of this dissertation includes data in New York. Two sets of empirical data sampled during the months of March, April, and May 2020, are included: (i) COVID-19 infectious cases and (ii) social media data.

First of all, Twitter APIs were used to collect tweets related to the coronavirus. Prior to developing the code for retrieving tweets, a few parameters needed to be determined: (1) the time window; (2) the targeted region; (3) the search keywords; and (4) the number of tweets.

To determine the time window, it was considered the initial peak of the COVID-19 outbreak when the WHO declared it a global pandemic on March 10, 2020. Around this time, national lockdowns (or quarantines) were implemented, and social distancing was recommended. Due to individuals maintaining distance and relying on social media for communication, it was valuable to examine data during this period.

In order to gain a deeper understanding of people's sentiments towards this unparalleled situation, the initiation of tweet data collection was set approximately ten days prior to the official declaration of the pandemic on March 1, 2020. The end time for tweets was decided based on reports concerning the conclusion of the stay-at-home orders in various states. Since the majority of states lifted the stay-at-home orders in middle to late May 2020, the designated end time for tweet collection was set to May 31, 2020.

In deciding on the target regions, I referred to the total number of infected cases in every state as of May 31, 2020, and ranked them in descending order. The infected

Table 3.1. Starting and ending dates of stay-home orders by state (as of May 31, 2020)

State	Starting date	Ending date
New York	March 22, 2020	May 15, 2020
New Jersey	March 21, 2020	June 9, 2020
Illinois	March 21, 2020	May 30, 2020
California	March 19, 2020	June 15, 2021
Michigan	March 24, 2020	June 1, 2020
Massachusetts	April 24, 2020	May 18, 2020
Pennsylvania	April 1, 2020	May 8, 2020
Texas	April 2, 2020	April 30, 2020
Florida	March 20, 2020	April 30, 2020
Maryland	March 30, 2020	May 15, 2020

and deceased cases as reported by CDC were employed. The numbers of infected cases and deaths caused by coronavirus in the top 10 states are described in Table 3.2 (CDC, 2021). New York City had the highest number of infected cases (206,857) and the highest death toll (22,131) in the United States of America. The state of New York had the second-highest number of infected cases (167,467) in the states, but was surpassed by the death toll in New Jersey. Therefore, the study scope in this research includes the infected- and deceased cases in both New York state and New York City.

3.2 Description of the Data

The COVID-19 data were collected from the CDC website (CDC, 2019) and the variables extracted are shown Table 3.3.

Social media data were collected from Twitter using Twitter API, a set of Python programming functions that enable access to Twitter feeds (Twitter, 2019). In alignment with the timeline for the collected COVID-19 cases, tweets that were created from March

Table 3.2. United States COVID-19 Cases and Deaths (%) by State (as of May 31, 2020)

State	Total Cases	Total Deaths
New York City	206,857	22,131 (11%)
New York	167,467	8,130 (5%)
New Jersey	160,807	11,698 (7%)
Illinois	120,260	5,390 (4%)
California	110,583	4,043 (4%)
Michigan	98,121	5,932 (6%)
Massachusetts	96,965	7,239 (7%)
Pennsylvania	75,940	5,567 (7%)
Texas	64,287	1,919 (3%)
Florida	55,131	2,644 (5%)
Maryland	52,778	2,653 (5%)
Total	1,209,196	74,693 (6%)

Table 3.3. Data specification of the collected COVID-19 cases

Description
States
The number of total cases
The number of new cases
The number of total deaths
The number of new deaths
Data created date and time

1, 2020 to May 31, 2020 were collected. Table 3.4 describes the specification of the collected tweets.

Table 3.4. Data specification of the collected tweets

Header
Posted dates
User ID
User name
Location
User description
Body texts
Total number of posts
Total likes
Total number of followers
Total number of followings
Account created date

3.2.1 COVID-19 cases

From March 1 to May 31, 2020, travel to the United States was restricted from more than two dozens European countries. Simultaneously, a nation-wide lock down and stay at home orders were issued across the states. Specific dates for the starting date and ending date of stay at home order in each state are provided in Table 3.1. For instance, the stay at home order in New York was started on March 22, 2020 and ended on May 15, 2020. Most of the states, excluding eight states (i.e., Iowa, Arkansas, North Dakota, South Dakota, Nebraska, Oklahoma, Wyoming, and Utah) started stay at home orders approximately in the middle of March and ending in the middle of May.

3.2.2 Twitter (X)

Twitter changed its name to X as of 2023 December, limiting access to the past tweets through API's. Thus, among the public domain, Chen et al. (2020a) published tweet ID's generated from posts related to COVID-19 using the list of keywords described in Table 3.5.

Table 3.5. The keywords used in searching tweets between March 1, 2020 and May 31, 2020 (Chen et al. (2020a))

#	Keyword	#	Keyword
1	Coronavirus	31	panic shop
2	Koronavirus	32	DuringMy14DayQuarantine
3	Corona	33	14DayQuarantine
4	covid-19	34	InMyQuarantineSurvivalKit
5	corona virus	35	coronakindness
6	sars-cov-2	36	quarantinelife
7	COVID 19	37	chinese virus
8	COVD	38	chinesevirus
9	pandemic	39	stayhome
10	coronapocalypse	40	stayhomechallenge
11	canceleverything	41	sflockdown
12	Coronials	42	DontBeASpreader
13	CDC	43	lockdown
14	Wuhancoronavirus	44	shelteringinplace
15	Wuhanlockdown	45	staysafestayhome
16	Ncov	46	saferathome
17	Wuhan	47	trumppandemic
18	N95	48	flattenthecurve
19	Kungflu	49	PPEshortage
20	Epidemic	50	GetMePPE
21	outbreak	51	covidiot
22	Sinophobia	52	covididiot
23	China	53	epitwitter
24	Sinophobia		
25	SocialDistancingNow		
26	Social Distancing		
27	SocialDistancing		
28	panicbuy		
29	panic buy		
30	panicbuying		

The shared tweet ID's were numeric digits which identified each tweet. In order to collect the source tweets with user name, body texts, etc., tweet ID's needed to be matched to the actual tweet post. However, if a tweet had already been removed before Twitter API's search and match process was executed, Python threw an error in the process of matching a tweet ID to its associated post. This error was handled by removing tweet ID's for all such deleted tweets.

An additional challenge was computational burden. The original tweet ID's shared by Chen et al. (2020a) included more than 300 million daily tweets related to COVID-19 worldwide. Thus, to minimize computational burden, tweets were randomly selected by generating random numbers for the index number of each tweet ID before filtering tweets only in New York. Thus, tweets from the state of New York were selected by considering the location field in a user bio. Note that, the state of New York including the New York City had the most infected cases as of 31 May, 2020. The state of New York was recorded by texts in several formats. Table 3.6 describes the search terms that were used to identify the state of New York.

Table 3.6. Search terms for the state of New York used in the python codes

State	Search terms
New York	, ny | new york | gotham | nyc | manhattan | long island | brooklyn | bronx | times square

3.3 Process for Retrieving the Data

3.3.1 Planning

In order to investigate the reasonable size of social media data for analysis, past studies using tweets in disaster or epidemics were referred. Chatfield and Reddick (2018) included 132,922 #sandy tweets from October 23 to November 10, 2012 to analyze the

interactions on social network during the time the Hurricane Sandy hit the Northeastern of the United States in late October of 2012. Cinelli et al. (2020) included 638,214 tweets from January 1 to February 14, 2020 to analyze how social media platforms handle questionable posts and credible posts. Lamsal (2021) selected 141,000 tweets worldwide with geographical coordinates in English from Jan 20 to April 18, 2020 for sentiment analysis. It has been observed that strikingly less number of tweets are geo-tagged. Burton et al. (2012) studied online health information and found that only about 2% of tweets were geo-tagged. Also, Qazi et al. (2020) studied that only 0.072% tweets of multiple languages in the context of coronavirus were geo-tagged. To improve such a small number of tweets with location, location information indicated by users is employed in this study.

3.3.2 Processing

For data retrieval, I used the tweepy library on python 3.8. All developments were conducted on my Samsung laptop (Intel Core i5 CPU, 8GB RAM). The initial data retrieval was conducted in May, 2022. Before retrieving data, it was important to check whether the tweet ID's were still available as of the early May 2022. However, the largest number of tweet ID's that can be queried at one time is merely 100 (i.e., batch size). Therefore, I randomly selected 100 unique id's to check their availability and repeated this process for 5,000 times (number of batches) for March, April, and May of 2020. Thus, the data could include a max of 500,000 tweets per month, if all the tweet ID's were available as of May 2022. Among available tweets, non-English tweets and all retweets were excluded. The remaining tweets were parsed into the following fields: date, tweet ID, body text, user ID, location, description, screen name, number of followers, number of followings, account created date, number of tweets the user liked during the account's lifetime, and total number of tweets (including retweets) issued by the user during the account's lifetime. This process took about five days.

Table 3.7 describes the number of tweets at each step of the retrieval process. For instance, after checking the availability of 1,500,000 tweet ID's, about 83% of these randomly sampled tweets were available (N = 3,755,752). Of those available tweets, about 39% tweets were in non-English reducing the data set further (N = 2,306,549). Removing the retweets (73%), the data set included 633,777 tweets, or about 38% of the English tweets. Finally, filtering for location New York yielded a final data set of 13,210 tweets. The entire process is shown in Table 3.7 and a sampled tweet is shown in Table 3.8. Tweet ID and user name were masked to avoid issues regarding privacy. Emojis, URLs, and special characters in the collected tweets were removed by using regular expressions. The number of the final tweets was not changed after cleaning. To verify the number of tweets, another data retrieval process was conducted in September, 2022. In this second run, the number of sample size is increased to 15,000 instead of 5,000 (conducted in May, 2022).

Table 3.7. The number of tweets (%) in the inclusion and the exclusion criteria (2020)

Description	March	April	May	Total
Random samples	1,500,000	1,500,000	1,500,000	4,500,000
Available	1,200,953 (80)	1,280,874 (85)	1,273,925 (85)	3,755,752 (83)
Non-English	430,326 (36)	521,338 (41)	497,539 (39)	1,449,203 (39)
English	770,627 (64)	759,536 (56)	776,386 (61)	2,306,549 (61)
Retweets	586,829 (76)	538,466 (71)	547,477 (71)	1,672,772 (73)
Original	183,798 (24)	221,070 (29)	228,909 (29)	633,777 (38)
& New York	4,442 (2)	4,227 (2)	4,541 (2)	13,210 (2)

Table 3.8. Example of the collected tweets

Field	**Value**
Date	2020-03-17 21:04:06
Tweet ID	1200011119999990000
Text	Trying to move during the coronavirus crisis under threat of the city entering a lockdown is one of the most stress
User name	johndoe1
Location	NYC
User description	writer
Account created date	2017-12-31 19:42:57
Followers	20
Total posts	223

4. PREDICTION OF INDIVIDUAL PREVENTION BEHAVIORS USING SOCIAL MEDIA DATA

The following summary is an excerpt a paper presented at the AI for Behavior Change Workshop of the 2023 Association for the Advancement of Artificial Intelligence (AAAI) Conference (Cho and Shehab, 2023).

4.1 Problem Description

Figure 4.1 presents a visual framework for how individual behaviors are influenced in a pandemic through social media. The structure of the framework is described as follows: in the COVID-19 outbreak, disaster relief organizations convey prevention guidelines to population. At the same time, people share their status via social media. Disaster relief teams leverage the quantitative analysis of prediction for individual prevention behaviors using artificial intelligence with social media data. Decision makers in disaster management organizations take actions after utilizing the quantitative analysis. Decision makers decide to promote the interventions with a probability of low compliance level by news and advertisements. Also, decision makers cooperate with social media platforms for providing personalized contents for individuals with low involvement in certain interventions. This study focuses on quantitative analysis by artificial intelligence and social media.

4.2 Data

In order to lower transmission, it is important to understand individual prevention practices for better intervention implementation. Individual preventive practices can be learned through social media data by machine learning. In particular, a self-training machine learning using tweets was developed and applied to predict individual prevention

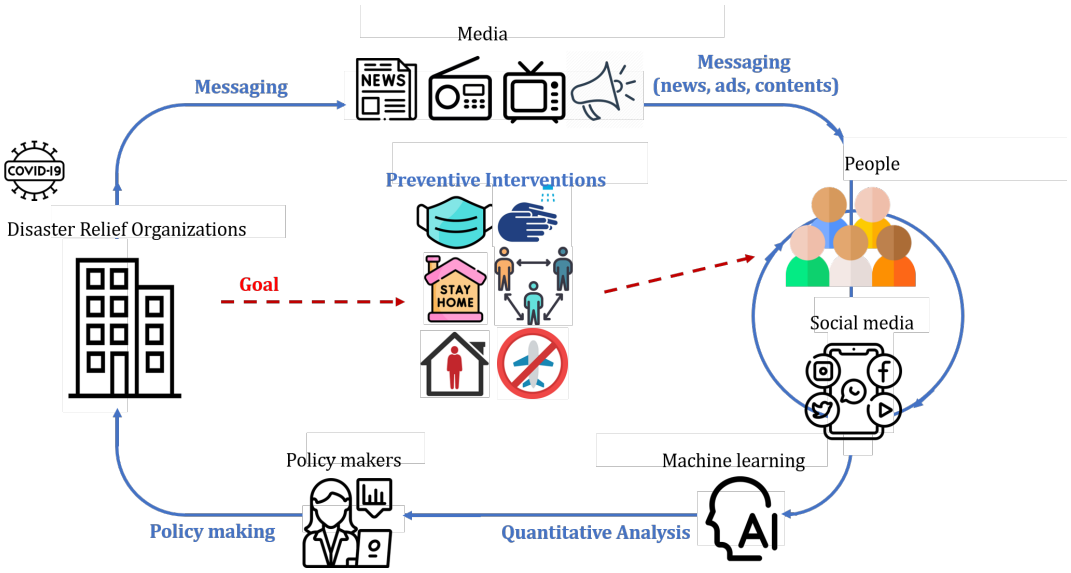


Figure 4.1. A framework of disaster management systems with prediction of individual prevention behaviors

behaviors in a disaster response phase. We collected tweets from March 1 to May 31, 2020 using Twitter API's. Due to the limited access to past tweets, we referred to publicly available tweet ID's (Chen et al., 2020a) and then identified their availability as of May 2022 by using the tweepy library in python. We refined the study scope to the state of New York, because it had the most infected cases in the United States of America between March 1 to May 31, 2020.

Figure 4.2 describes data retrieval process. Due to the computational burden, we randomly sampled 500,000 tweets per March, April, and May for checking tweets' availability. We verified whether the number of tweets in the prevention behaviors were balanced. If not, we collected more tweets for underrepresented behaviors. The progression of these steps with Intel Core i3 CPU, 2.30 GHz, 8.00 GB RAM necessitated approximately a week's duration for completion. We filtered out non-English tweets by checking "lang=en" of each tweet's user object and removed re-tweets by searching "RT @" in body text. The number of tweets in English and without retweets was 163,821

worldwide. Among them, tweets with location information in user bios are selected if their "location" of user objects matches New York.

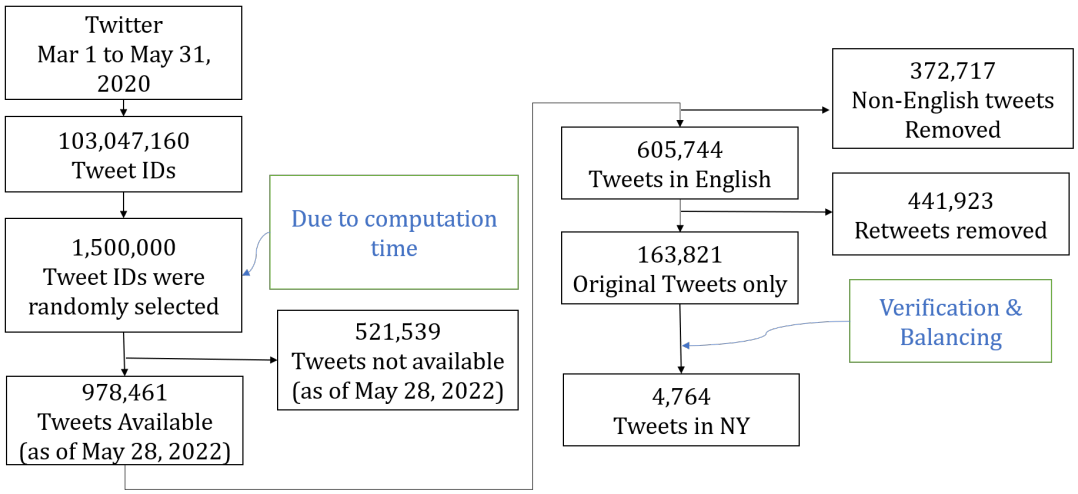


Figure 4.2. Inclusion and exclusion criteria of tweets retrieval

Each data entry included the tweet's created date, tweet ID, main text (or body text), user name, location, user description (or user bio), account created date, number of followers/followings, total number of tweets a user has liked during the account's lifetime, and total number of tweets (including re-tweets) created by a user. Table 4.1 describes examples from the collected tweets. We replaced tweet IDs and user-specific information with mock data.

4.2.1 Data preprocessing

Emojis, urls, special characters were removed by regular expressions. Stop words in each tweet were removed if they were included in the nltk corpus. We verified that each tweet was created in March 1 to May 31, 2020 by the data filtering function in Excel. Additionally, body texts collected in data were randomly verified whether they had relevant keywords to coronavirus in Excel. All numeric features of tweets such as the

Table 4.1. Examples of the collected tweets (Tweet ID's and user names are hidden)

Header	**Data**
Date	2020-03-17 21:04:06
Tweet ID	*1200011119999990000*
Text	Trying to move during the coronavirus crisis under threat of the city entering a lockdown is one of the most stress
User name	johndoe
Location	NYC
User description	writer of things, mostly comics
Followers	20
Followings	11
Total likes	250
Total posts	223
Account created date	2018-05-12 23:13:54
Date	2020-05-15 14:27:55
Tweet ID	*1200011119999990001*
Text	My job is requiring covid antibody test. Soon I will know if I am immune. If I am immune, I will be unstoppable
User name	janedoe
Location	New York
User description	someday everything is gonna be different when I paint that masterpiece
Followers	279
Followings	1362
Total likes	2464
Total posts	3153
Account created date	2018-01-23 17:26:27

number of followings/followers, number of likes, and number of posts were standardized by log function. Outliers were removed. Polarity scores and subjective scores were evaluated based on given data. TextBlob library in python was used. These scores were included in a feature set to augment the features for better prediction of the individual prevention behaviors. Figure 4.3 shows the word cloud of collected 4,764 tweets in the state of New York. These tweets included the preventive behaviors of relevant terms.

Figure 4.3. Word cloud of collected tweets in the state of New York

4.2.2 Encoding (prevention behaviors lexicon)

If tweets included appropriate keywords for certain prevention, they were then labeled by that prevention behavior. Table 4.2 provides an overview of the preventive behaviors and associated conceptual terms (or keywords) utilized in encoding tweets prior to model training. Equivalent conceptual terms corresponding to each preventive behavior were collected from sources including the CDC, scholarly literature, and online repositories (CDC, 2020; Kwon et al., 2020; Wikipedia, 2021). These categories encompass: mask usage; hygiene practices; social distancing protocols; quarantine measures; and travel restrictions. The validation of the lexicon involved the augmentation of additional keywords and subsequent reassessment through the identification of absent synonyms for the prevention behaviors using the word cloud. Any identified missing synonyms were subsequently incorporated into the lexicon.

In cases where a tweet encompassed multiple prevention behaviors, preference was given to the prevention method mentioned most frequently. In situations where multiple prevention behaviors were equally mentioned within the same tweet, priority was allocated based on the level of restrictiveness, adhering to the sequence: travel ban, quarantine, distancing, mask, hygiene.

Table 4.2. Interventions (prevention behaviors) by CDC

Prevention	Relevant words
Mask	mask(s), facemask, wearamask, face coverings, face shields, respirator(s), N-95 (n95, nn95), KF-94 (kf94), ppe

continued on next page

Table 4.2. *continued*

Prevention	**Relevant words**
Hygiene	hygiene, hygienic care, hygienic, hand washing, wash(ing) (your) hands, cover(ing) coughs and sneeze, sanitize(d), sanitizer, disinfect, disinfection, disinfectant(s), avoid poorly ventilated, avoid closed spaces, intervention(s), prevention(s), preventive, guideline(s), guidance(s)
Distancing	physical distancing, social distancing, social distance, distancing, distance, 6 feet, flatten the curve, keep(ing) space, give(ing) space, no (large) gathering, no party, avoid crowds
Quarantine	quarantine(s), self-quarantine, isolation, curfew, separate(s) and restrict(s) the movement(s), business closures, lockdown, lock down, shutdown, stay(ing) home, stay-at-home, stay at home
Travel ban	travel ban, travel(ing), travel(ed), tour, tourism, fly(ing), cruise, ship, private jets, train, transportation, border closures, travel restrictions, cancel trip, flying jet, flight(s), airplane(s), airbus, airline(s)

Basic hygienic behaviors such as hand washing and disinfecting were categorized into hygiene. If tweets included [prevention], [intervention], or [guidelines], but did not explicitly include the lexicon, then these tweets were grouped into hygiene.

This encoding process only considered whether input tweets included the keywords specified in the dictionary. After extracting the keywords, only less than 15% tweets explicitly contained the keywords. These tweets were manually validated by inspecting whether each context conformed with the given label, or prevention behavior.

4.3 Prediction of Individual Prevention Behaviors

The question of predicting adoption of intervention behaviors require predicting individual prevention behaviors. The most likely preventive behaviors can be inferred by probabilities given social media data. Processing social media data involves the reduction of textual information into a numeric vector space, achieved through pre-processing, encoding, and TF-IDF vectorizer. Given that the labeled dataset constitutes a small fraction of the overall collected data, the present study engaged in semi-supervised learning. To address the problem, a machine learning with self-training scheme is developed (Zoph et al., 2020; Baevski and Mohamed, 2020). Figure 4.4 conceptualizes the methods, `PRecomm`. In the initial training, the labeled tweets were partitioned into training and testing sets to facilitate classical machine learning classifiers. Specifically, this research employed random forest, support vector machine, and k-nearest neighbor algorithms for analysis. After the initial training, PRecomm was trained with both original training data and data with predicted labels (or pseudo-labeled data) from the pre-trained PRecomm. The new-trained PRecomm makes a new prediction for unlabeled data. This process is defined as self-training. This self-training process was repeated until the stopping criteria was met. A heuristic approach was used for stopping criteria, in which if improvement in F-1 score was not greater than the previous iteration, then iteration was stopped. Pseudo code of `PRecomm` is shown in Algorithm 1 in Appendix A2.

Random forest was used as a baseline classifier as compared to support vector machines. The candidates for the number of estimators for random forest were {1, 2, 8, 32, 90, 100, 110}. Support vector machines with linear kernel was used to predict the most

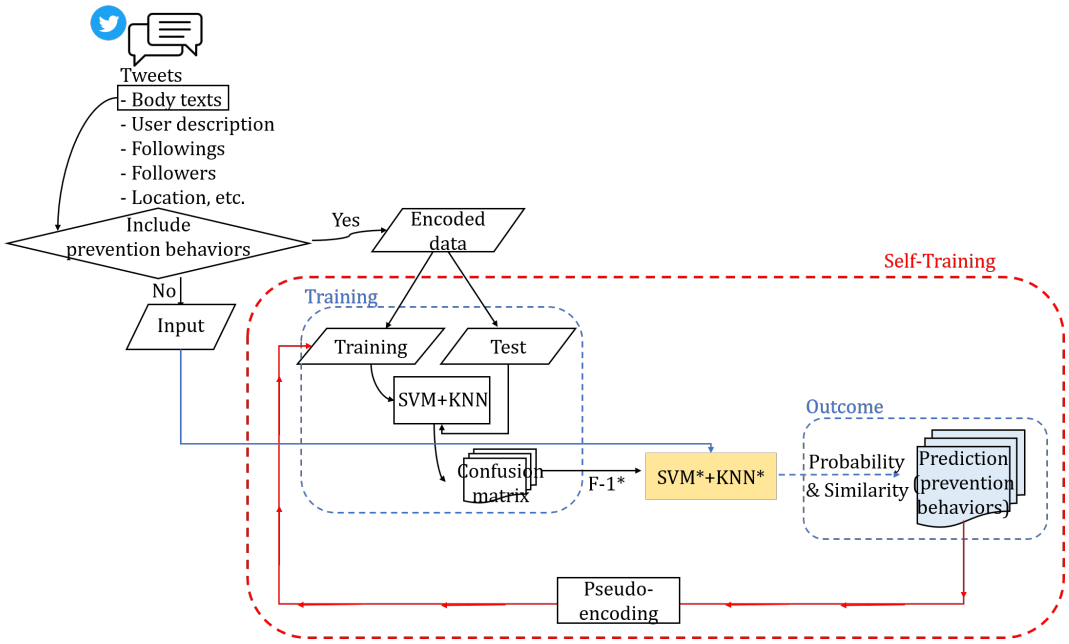


Figure 4.4. Prediction and Recommended Predictions (PRecomm)

likely prevention behavior of each input tweet. The candidates for regularization parameter (C) were {0.1, 0.2, 0.7, 0.9, 1.0 ,1.8, 2.0}. K-nearest neighbor was used for prediction of multiple preventions. It was assumed that close tweets share similar labels. PRecomm ranks the nearest tweet ID's from k-nearest neighbor in descending order of similarity scores (i.e., `1-distance`). Top two prediction results with the highest frequency were assigned to input tweets. Cosine distance was used, because it gives closer distance than ℓ_2 norm. The candidates for the number of neighbors (K) were {1, 2, 9, 10, 12, 19, 20}. Labeled and pseudo-labeled tweets were split into training and testing with the ratio of 80% to 20%. 5 fold cross-validation was used to both estimate the parameters of classifiers and subsequently adjust classifiers to provide more accurate probability of prediction. Model performance was measured by F-1 score and efficiency (or computation time). Prediction probabilities for unlabeled data by support vector machines and

random forest was evaluated using `predict_proba` from the sklearn package. Similarity scores for unlabeled data by k-nearest neighbor was evaluated using `1-distance`.

4.4 Results and Discussion

4.4.1 Tweets results

Only 541 (13%) tweets contained the prevention behaviors lexicon. We collected more tweets for underrepresented prevention in the encoded data such as mask, distancing, and travel ban. Consequently, 663 tweets (14%) were included as the encoded tweets after balancing and used for training the model. The 4,101 (86%) tweets were unlabeled and predicted by the PRecomm. Table 4.3 describes the distribution of labeled and unlabeled tweets from New York in each month. All tweets were unique. We defined noise in tweets when users posted tweets multiple times with specific prevention behaviors. Seven tweets were determined to be noise and removed from study data. 663 unique tweets from 631 unique users were included in this study.

Table 4.3. The number of tweets of labeled tweets, labeled tweets after balancing, unlabeled tweets, and total tweets in the state of New York

New York	**March**	**April**	**May**	**Total**
Labeled	131	194	216	541
Balanced	216	216	231	663 (14%)
Unlabeled	1,378	1,345	1,378	4,101 (86%)
Total	1,594	1,561	1,609	4,764(100%)

Figure 4.5 describes the prevention behaviors distribution before and after balancing. Table 4.4 also describes the number of tweets in each prevention before and after balancing. The disproportion inherent in the original random sampling was partially considered in balancing.

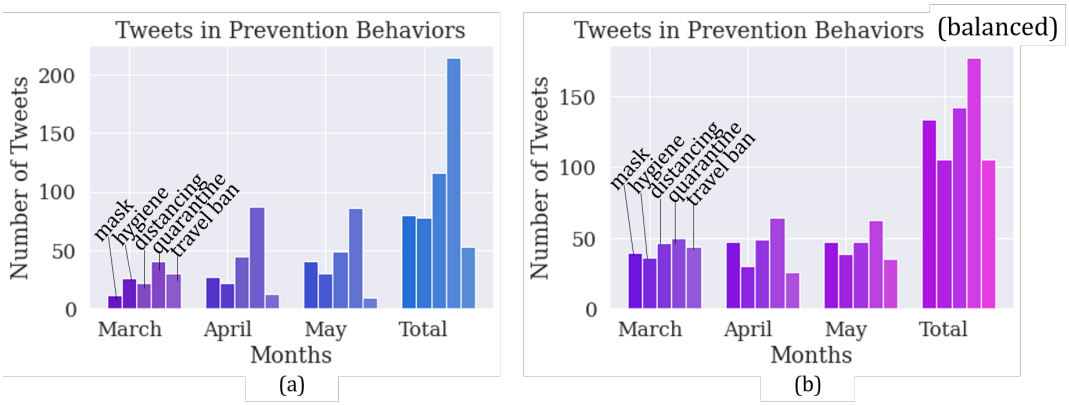


Figure 4.5. Original (a) and balanced (b) prevention behaviors distribution

Table 4.4. The number of tweets in each prevention behavior after balancing (the number of tweets before balancing is described in parenthesis)

Prevention	**March**	**April**	**May**	**Total**
Mask	40 (12)	47 (27)	47 (41)	134 (80)
Hygiene	36 (26)	30 (22)	39 (30)	105 (78)
Distancing	46 (22)	49 (45)	47 (49)	142 (116)
Quarantine	50 (41)	64 (87)	63 (86)	177 (214)
Travel ban	44 (30)	26 (13)	35 (10)	105 (53)
Total	216 (131)	216 (194)	231 (216)	663 (541)

We observed 30 (5%) tweets with multiple prevention behaviors when encoding. These tweets were encoded by a single prevention based on the decision rule of frequency and restrictiveness discussed in Section 4.4.

4.4.2 Training results

Number of iteration of self-training

For the baseline classifier and support vector machines, two iterations of self-training improved F-1 scores the most, whereas for k-nearest neighbor, single self-training improved F-1 score the most. Figure 4.6 describes the F-1 scores of hyper-parameter tuning for the baseline classifier (i.e., random forest), support vector machines, and k-nearest neighbor in the initial training and at each iteration of self-training.

Hyper-parameters

After hyper-parameter tuning, We choose the regularization parameter (C) for support vector machines as 0.7 with F-1 score of 0.98. For random forest, the number of estimators was selected as 100 (F-1=0.74). We choose the number of neighbors for k-nearest neighbor as $K = 10$ (F-1=0.55).

4.4.3 Model evaluation results

Baseline vs support vector machines

F-1 score of support vector machines was evaluated as 0.97, which is 31% higher than F-1 score (F-1 = 0.74) of the baseline classifier (random forest). The computation time of the baseline classifier was 6,618 seconds to complete initial and three iterations of self-training, which was 262% more expensive than support vector machines. Random forest had an average prediction score of 0.72, which is similar to the support vector

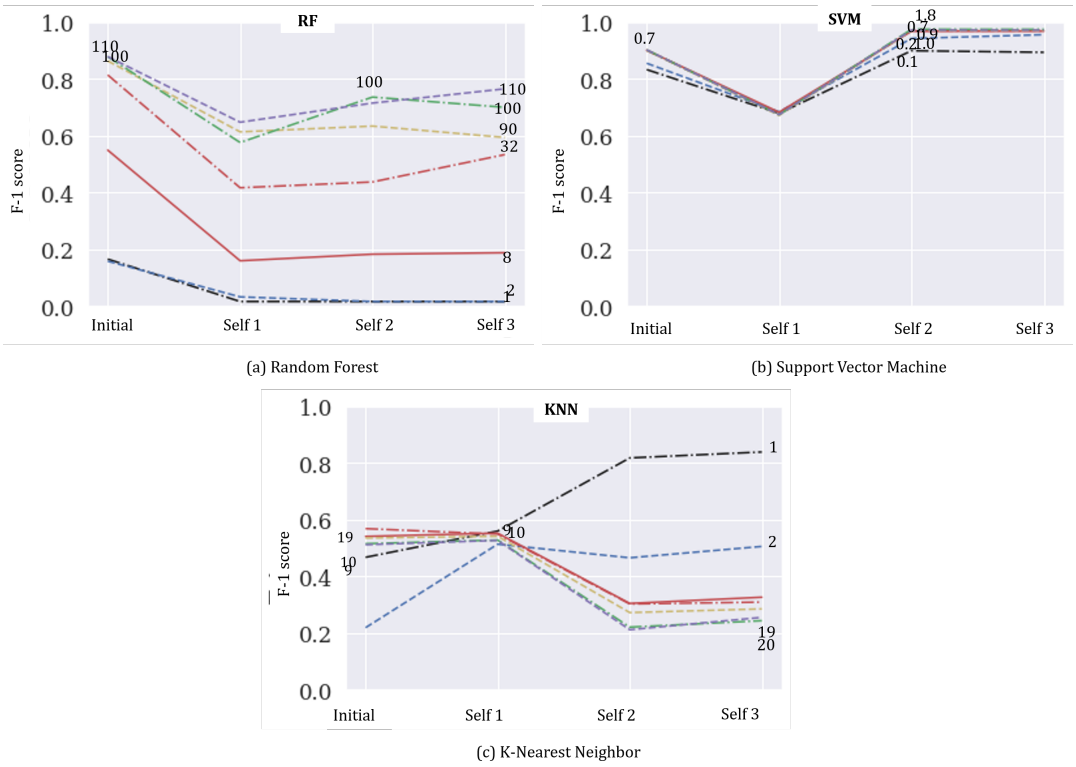


Figure 4.6. F-1 scores for hyper-parameters tuning (a) random forest, (b) support vector machines, and (c) k-nearest neighbor in the initial training and each iteration of self-training

machines' average prediction score of 0.73. Therefore, support vector machines were selected instead of the baseline classifier in consideration of both accuracy and efficiency.

K-nearest neighbor

While support vector machines achieve the highest F-1 score for predicting the most likely prevention, k-nearest neighbor provides the fastest computation for multiple preventions. K-nearest neighbor had accuracy score of 0.64 and F-1 score of 0.55. Its computation time was 1,172 seconds for completing self-training including hyperparameter

tuning and ordering of multiple prevention behaviors by similarity scores. Its average similarity score of prediction results was 0.8.

Table 4.5 describes accuracy, F-1, computational time, and prediction scores or similarity scores of random forest (RF), support vector machines (SVM), and k-nearest neighbor (KNN). Figure 4.7 describes F-1 scores across the different classifiers in the initial- and self-training.

Table 4.5. Model performance metrics of random forest (RF), support vector machines (SVM), k-nearest neighbor (KNN)

Metrics	**RF**	**SVM**	**KNN**
Accuracy	0.84	0.99	0.64
F-1	0.74	0.97	0.55
Time (seconds)	6,618	1,827	1,172
Prediction score	0.72	0.73	-
Similarity score	-	-	0.80

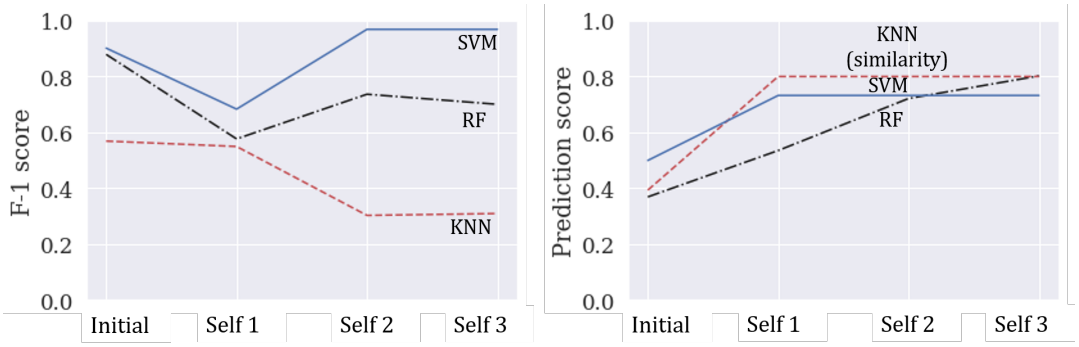


Figure 4.7. F-1 scores and prediction probabilities or similarity scores by random forest, support vector machines, and k-nearest neighbor

4.4.4 Prediction results

Individual tweets

Table 4.6 describes one example of the results from the PRecomm. PRecomm predicted the input tweet as travel ban with probability of 0.91. Also, it listed 10 similar tweets with similarity scores for any chance of multiple prevention behaviors it might had, then it predicted the input tweet as hygiene, travel ban, and mask based on the frequency. Similarly, the rest of 4,101 unlabeled tweets were predicted in the same format as Table 4.6.

Aggregated prediction results

Prevention behaviors distribution after prediction is as follows: hygiene (1221 (30%)), travel ban (1094 (27%)), quarantine (1067 (25%)), mask (524 (13%)), distancing (195 (5%)). Also, average prediction probabilities across different classes was 0.73 with standard deviation of 0.02. Distribution of prediction probabilities across prevention behaviors are described in Figure 4.8.

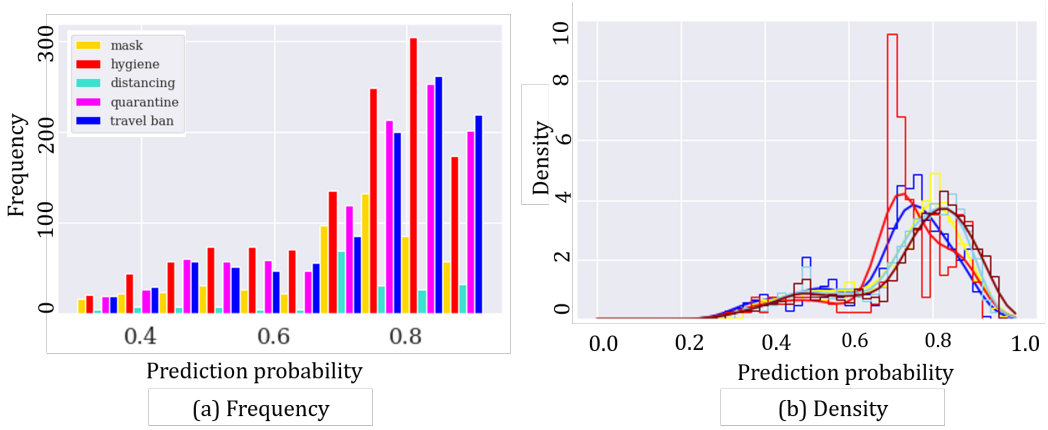


Figure 4.8. Prediction scores distribution - SVM

The prediction probabilities were improved with self-training machine learning, as the average prediction score became higher than the one without self-training machine learning (Figure 4.9)

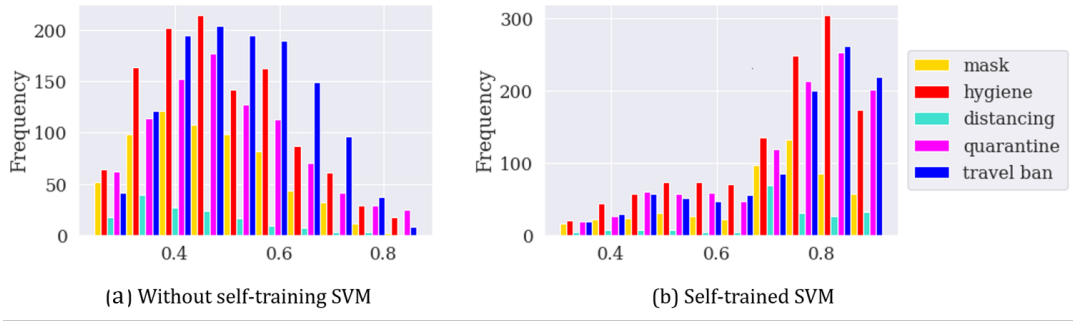


Figure 4.9. Prediction scores distribution without (a) and with self-training (b)

4.4.5 Discussion of the Prediction of Individual Prevention Behaviors

Self-training performance was inherited by selection of machine learning classifiers. The parameters for classifiers after cross validation without or with self-training were not changed. Even if any change in F-1 score existed, it was that the original parameter still had higher F-1 score than the other candidates after self-training. In addition, it was observed that self-training with random forest (with the number of estimators = 100) converged to a single class even after three iterations where all pseudo labeled data was inputs to the model at once at each iteration. The ratio of original training data to pseudo-labeled data was experimented using 90% to 10% (and 70% to 30%) iteratively. The pseudo-labeled data was randomly selected at each iteration. However, the small portion method did not fix the issue of convergence to a single class but just took more time to reach to the convergence. Therefore, it seems that small imbalance in classes can cause self-training machine learning to converge to the majority class regardless of the ratio of pseudo labels. Hence, whole pseudo labels was leveraged at once in self-training

and the and observed F-1 scores at each iteration in deciding stopping timing. Figure 4.10 describes confusion matrices of self-training with the small portion approach.

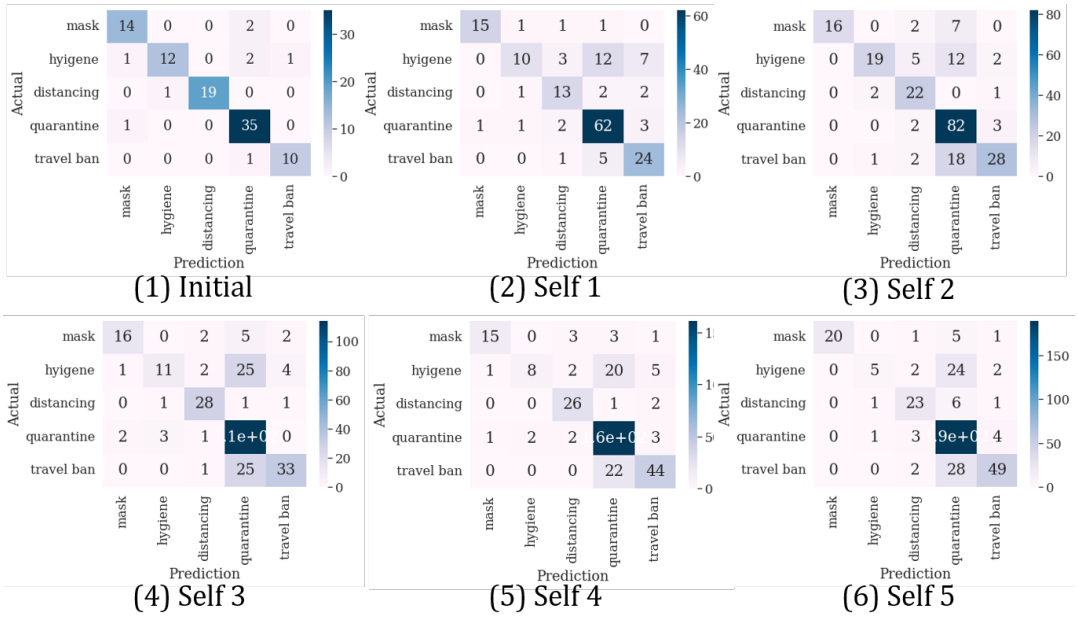


Figure 4.10. Confusion matrix at each iteration of self-training with random forest (a ratio of training data to pseudo-labeled data was 90%:10% at each iteration)

Although self-training did not seem to significantly improve model performance (F-1 score), it was observed that both average prediction probabilities in support vector machine and random forest and average similarity scores in k-nearest neighbor were improved after self-training. Even though training without self-training seemingly had good model performance in our experiment, it was partially due to well encoded and balanced training data. It required meticulous work in construction of the lexicon for encoding and re-collection of data for balancing. Therefore, it is recommendable to experiment how self-training performs in the situation where such sedulous work is not available.

In consideration of efficiency and accuracy, prediction made by support vector machines is preferable to the baseline classifier in predicting the most likely prevention behavior. It is not negligible that the baseline classifier took 3.6 times slower than support vector machines in contemplating the fact that tweets size exponentially grows in practice. K-nearest neighbor was able to predict multiple prevention behaviors in a short amount of time. Therefore, k-nearest neighbor results can be adopted along with support vector machines results for better flexibility in prediction.

From perspective of practical application of the aggregated results of preventions, about 25-30% of individuals were likely to follow the prevention of hygiene, travel ban, and quarantine, whereas only 5-13% of population were likely to follow mask and distancing. Therefore, while sustaining high compliance level with hygiene, travel ban, and quarantine, it is recommended for disaster management organizations to concentrate on deploying mask and distancing as well. It is noteworthy that the utilization of this quantitative measure of individual compliance with intervention has the potential to improve a mitigation plan to prevent transmission. Policy makers can promote interventions with low compliance level, deploying personal protective tools in order to improve the compliance level. Further, the compliance level with intervention can be further investigated based on location, providing local policy makers the quantitative evidence for updating the prevention strategies. As our PRecomm can evaluate the individual compliance level with intervention in real-time, policy makers and disaster relief organization can reduce response time and improve the efficacy of prevention strategies during emergencies. Moreover, the implementation of traditional methodologies such as observation or survey studies during emergencies is frequently challenged by cost constraints. Hence, the utilization of self-training machine learning techniques to quantify individual compliance levels with preventive measures using social media data offers a viable alternative for policy making during emergency situations where conventional approaches are not feasible.

Table 4.6. An example of the PRecomm outcome

Field	**Value**
Time	2020-03-01 23:39:11
Tweet ID	1200008100001990001
Body	Diamond Princess passenger: I have the coronavirus. So far, it isn't that bad. via Advocate
Location	Manhattan
User description	Father husband volunteer fireman data analyst. Occasionally life gives us fairytale
User name	jacobdoe
#Followers	3.60
#Followings	2.63
Account created date	2014-02-14 23:28:02
#Likes	3.88
# Posts	4.31
Predicted prevention	***travel ban***
Probability score	***0.91***
Nearest neighbors: score	hygiene: 1.0, distancing: 0.1, travel ban: 0.09, mask: 0.09, hygiene: 0.09, mask: 0.08, mask: 0.08, travel ban: 0.08, hygiene: 0.07, travel ban: 0.07
Prediction (frequency)	hygiene(3), travel ban(3), mask(3), distancing(1)

5. INCORPORATION OF THE EFFECT OF PREVENTION BEHAVIORS WITH EPIDEMIC MODELING

This manuscript has been submitted to the Institute of Industrial and Systems Engineers (IISE) Transactions on Healthcare Systems Engineering in September 2023 and is in revision.

5.1 Problem Formulation

Classical epidemic modeling method, compartmental differential equations, was employed in this study. In this model, there are five states, susceptible (S), exposed (or latent) (E), infected (I), and recovered (R) or deceased (D). Figure 5.1 illustrates the states and associated parameters that govern the rate of population change as it transitions from one state to another.

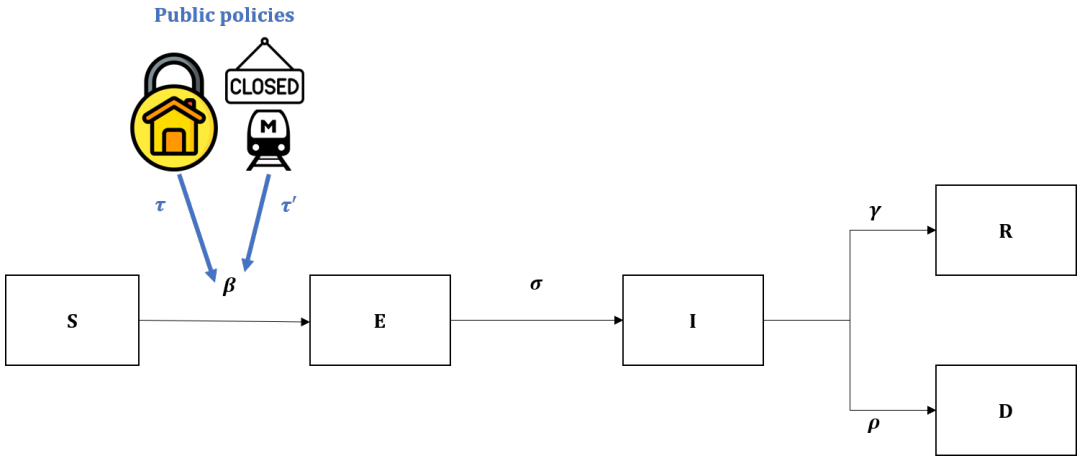


Figure 5.1. The SEIRD model with nonpharmaceutical interventions and model parameters

Nonpharmaceutical interventions (NPIs) aim to reduce the transmission rate between individuals in the susceptible and exposed populations by a factor of τ. Once patients transition to the recovered or deceased state, they are assumed not to re-enter the SEIRD

system. The changes in population size in each state are mathematically formulated in Equation 5.1.

$$\begin{aligned}
\frac{dS}{dt} &= -\tau\beta(t)SI/N \\
\frac{dE}{dt} &= \tau\beta(t)SI/N - \sigma E \\
\frac{dI}{dt} &= \sigma E/N - (\gamma + \rho)I \\
\frac{dR}{dt} &= \gamma I \\
\frac{dD}{dt} &= \rho I.
\end{aligned} \tag{5.1}$$

N denotes the total number of population such that $N = S(t)+E(t)+I(t)+R(t)+D(t)$, *S(t)* denotes the number of susceptible population at time t who are not infected yet, *E(t)* denotes the number of population exposed to infected population but not infectious yet at time t, *I(t)* denotes the number of population who are infected at time t, *R(t)* denotes the number of population who get recovered at time t, and *D(t)* denotes the number of deceased population at time t. $\beta(t)$ is a transmission rate at time t, τ is the effect of nonpharmaceutical intervention, σ is an exposed rate, γ is a recovery rate, and ρ is a mortality rate. N is assumed to be a static variable, while the transmission rate $\beta(t)$ is assumed to be a dynamic variable in the course of the disease. Details will be discussed in Section 4.3. Intuitively, $\beta(t)$ is interpreted as a probability of transmission for a single individual after contacting with the potential infectious population ($\frac{I}{N}$). Table 5.1 describes the parameters and notations in the model and their definition.

Table 5.1. Notation

Notation	Definitions
β	Rate from susceptible to exposed
σ	Rate from exposed to infectious
γ	Rate from infected to recovered
ρ	Rate from infected to deceased
τ	NPI effect on March 23
τ'	NPI effect on April 30

A general representation of the parameter estimation problem can be formulated as Equation 5.2, where the objective is to minimize the mean squared errors between the reported infected cases and the estimates of infected cases. The observation y represents the reported infected cases on day t, and the goal is to find parameter values that result in the minimum squared residuals between the model's predicted values and the observed data.

$$\begin{aligned} \arg\min_{\theta} \quad & \frac{1}{T}\sum_{t=1}^{T} \|y(t) - h_\theta(t)\|^2 \\ \text{s.t.} \quad & \theta \in \Theta, \end{aligned} \tag{5.2}$$

where h_θ is our hypothesized function, which is the integral of the rate of change of infected cases (dI/dt) as described in Equation 5.1. The parameter vector θ represents the values of various parameters in the model, such as $\theta = [\beta(t), \sigma, \gamma, \rho, \tau, \tau']^T$, where t represents time in days and Θ is a feasible set of θ. Each parameter value, such as exposed-, recovered-, and death- rate, are bounded within a tighter range of values in reality.

5.2 Data & Experiment Settings

5.2.1 Data

This study utilized the infected and deceased counts data reported by the CDC for the state of New York during the period from March 1 to June 8, 2020 (CDC, 2020). Simulation method was employed to study the dynamics of the coronavirus outbreak by tracking the number of individuals in different compartments, Susceptible (S), Exposed (E), Infected (I), Recovered (R), and Deceased (D), over a hundred-day period that included the first peak of the coronavirus outbreak. The study also considered the effect of nonpharmaceutical interventions, accounting for the realistic timeline of public policies implemented in New York. The total population estimate of New York was 19,378,102 (N) on April 1 2020 (United States Census, 2020).

5.2.2 Public policies - interventions

Before vaccines were developed (and even after vaccines became available) during the COVID-19 pandemic, public policies were implemented in affected regions for decreasing respiratory infections and reducing the contact rate per individual, to intervene the transmission chain. Table 5.2 summarizes the timeline of public policies implemented in New York state in March 2020. The New York governor banned gatherings of more than 500 people on March 12 (NewYork, 2020). Local schools temporarily shut down on March 13, followed by NYC closed public schools on March 15. On March 23, all schools and businesses were closed (i.e., New York on pause). Then, on April 30, the subway in New York City stopped the service from 1-5 AM for disinfection (i.e., metro shutdown).

In this study, it was assumed that nonpharmaceutical interventions aimed at altering the transmission rate began on Saturday, March 23, when New York state went on pause and all non-essential businesses were closed (i.e., $\beta(t) = \tau\beta(t-1)$). Then, it was assumed

that another intervention of metro closure was implemented on April 30. Similarly, we divided the time horizon into smaller segments, aligned with the timeline of events, and estimated the $\beta(t)$ for each time segment.

Table 5.2. COVID-19 Timeline in New York

Date	Orders
March 12	Banned gatherings of more than 500 people
March 13	Local Schools shut down
March 15	Closed public schools in NYC
March 23	All non-essential for profit or non-profit businesses statewide to close their in-office personnel functions (New York on Pause)
April 30	NYC subway closures from 1 a.m. to 5 a.m. during the coronavirus pandemic in order to disinfect trains and stations (metro shutdown)
May 20	Daily new hospitalizations drop below 5,000 daily for the first time

5.2.3 Estimation of transmission rate - deterministic dynamic variable

In the simulation model, we assumed that the transmission rate β is not static, instead dynamic; β changes over time, particularly it decreases when public policies are implemented. Figure 5.2 illustrates this dynamics of the transmission rate with a discontinuous step function. Modeling with a dynamic variable of the transmission rate allows for a more accurate estimation of the time-varying evolution of the COVID-19 outbreak.

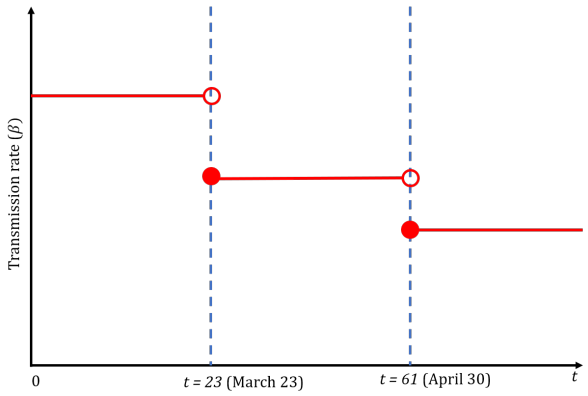


Figure 5.2. A theoretical step function of the dynamic variable of transmission rate assuming a public policy is implemented on March 23 and another public policy is implemented on April 30

In our model, we update the value of β at time t was updated based on the starting times of COVID-19 interventions described in Table 5.2 ($t = 0$ indicates March 1 and $t = 100$ indicates June 8, 2020). In this study, we considered two major interventions in the modeling process, which includes the New York lockdown on March 23 and the metro shutdown on April 30. The decreasing rate of the transmission rate due to public interventions is denoted by τ's such that $0 < \tau < 1$. τ denotes the decreasing rate of transmission rate by the New York on pause, while τ' denotes the decreasing rate of transmission rate by the metro shutdown. This can be formulated by:

$$\beta'(t) = \tau\beta(t-1), \qquad \text{for } \tau \in (0,1) \text{ and } t = 23.$$

The updated transmission rate remains unchanged until the next update. Similarly, on April 30, the transmission rate was further reduced due to another intervention of metro

closures in New York City. In this study, we assumed that the effectiveness of metro closures results in a further decrease in the transmission rate by τ' such that:

$$\tilde{\beta}'(t) = \tau'\beta'(t-1), \qquad \text{for } \tau' \in (0,1) \text{ and } t = 61.$$

5.3 Methods

5.3.1 Grid search method

In the grid search approach, different combinations of values for the parameters β, σ, γ, ρ, and τ's are plugged into Equation 5.1 and 5.2. Table 5.3 describes the ranges of values for each parameter used in this study based on CDC reports (CDC, 2022). Then, the mean squared error was calculated for 3.24 million combinations resulted from multiplying parameters' dimensions (i.e., 3.24 million= $||\beta \cdot \sigma \cdot \gamma \cdot \rho \cdot \tau \cdot \tau'||$). The step size was determined empirically by examining a range of parameter values and the number of iterations. That is, the step size was evaluated by dividing a range of parameter values by the number of iterations. This approach allowed for a systematic exploration of parameter space to find the combination of values that minimizes the mean squared errors.

Table 5.3. Range of parameter values for grid search

Parameter	Range	Dimension	Incremental value
β	[0.2, 0.9]	$\lvert\beta\rvert \leq 20$	0.0335
σ	[0.7×10^{-1}, 0.5]	$\lvert\sigma\rvert \leq 20$	0.0215
γ	[0.7×10^{-1}, 0.2]	$\lvert\gamma\rvert \leq 10$	0.0130
ρ	[0.1×10^{-3}, 9.2×10^{-2}]	$\lvert\rho\rvert \leq 10$	0.0092
τ	[0.5×10^{-1}, 9.5×10^{-1}]	$\lvert\tau\rvert \leq 9$	0.1000
τ‘	[0.5×10^{-1}, 9.5×10^{-1}]	$\lvert\tau$‘$\rvert \leq 9$	0.0900

5.3.2 Trust region method

To justify the selection of trust region interior point, we illustrate the preliminary result from three methods including sequential least squares programming (SLSQP), linear approximation (COBYLA), and the trust region interior point method (trust region) in Figure 5.3. 5.4 compares the details of result. The mean squared error exhibited a descending trend in the following sequence: SLSQP, COBYLA, and trust region.

To solve Equation 5.2, we applied nonlinear optimization algorithms from the optimize package in the Scipy library in python. After investigating three different nonlinear optimization methods, the trust region method was selected in this study based on the fitting result with the least mean squared error.

The primary goal of this study was not to discover a nonlinear solver or demonstrate algorithmic effectiveness. Instead, it focused on solving the optimization problem to analyze the evolution of COVID-19 and assess the effectiveness of nonpharmaceutical interventions. As a result, delving into the specifics of each algorithm and its performance was beyond the scope of our research. Instead, we compared the mean squared errors by 3 different methods in the preliminary study. We investigated the trust region interior point method had the best objective value among 3 methods for finding parameters describing the COVID-19 dynamics. Therefore, the trust region interior point method was selected as a nonlinear optimization solver in this study.

Barrier function in trust region interior point

In applying nonlinear optimization algorithms, Equation 5.2, which is a constrained problem, can be transformed to a unconstrained problem shown in Equation 5.3 by introducing a penalty term, or a log-barrier term, consisting of each constraint to the original objective function. For instance, if θ is greater than 0 and less than 1, where θ

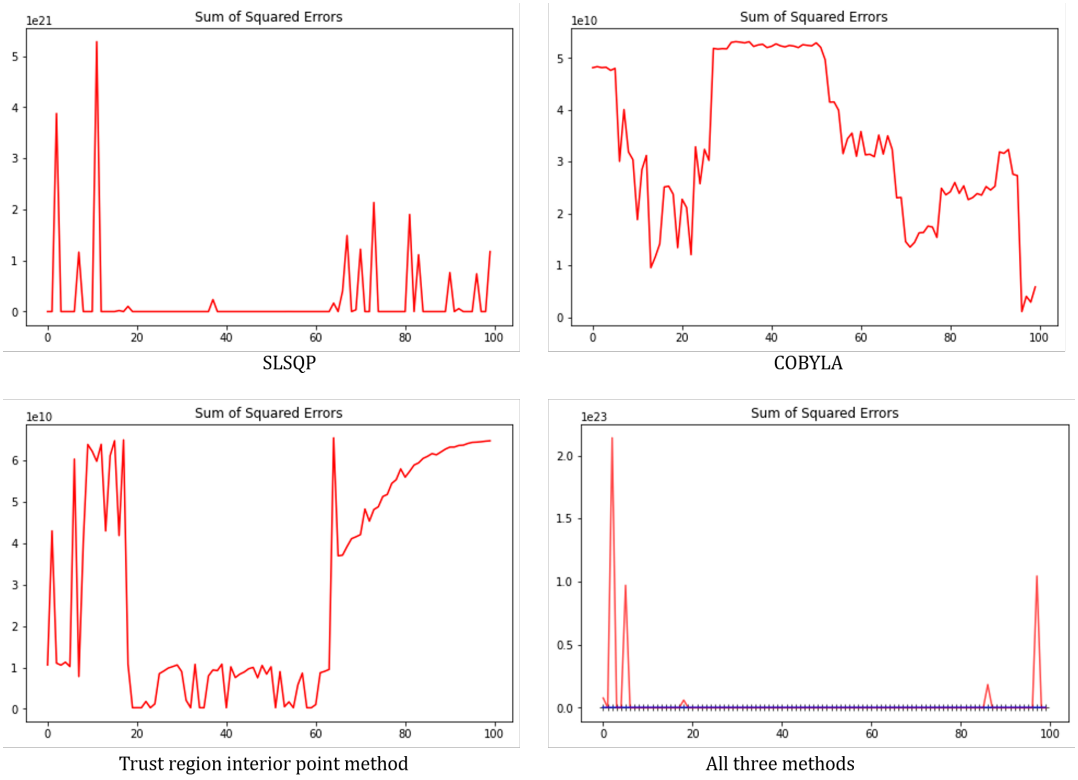


Figure 5.3. MSE by SLSQP, COBYLA, Trust region interior point, and a combined graph of results from all three methods

is a parameter in Θ in Equation 5.2, then the penalty term $-\mu(\log\theta+\log(1-\theta))$, where μ is a barrier parameter such that $\mu > 0$ is added to the original objective function.

$$\arg\min_{\theta} \quad \frac{1}{2}\sum_{t=1}^{T}\left\{\|y(t)-h_\theta(t)\|^2\right\} - \mu(\ln(-\theta+1)+\ln\theta), \tag{5.3}$$

Convergence of trust region interior point

The trust-region interior-point method defines a neighborhood, known as the trust region or the trust region radius (r), around the current solution based on the performance in the previous iteration, and then searching for the direction and step size within that neighborhood that approximates the minimizer of the problem (Nocedal and Wright,

Table 5.4. Comparison of the mean squared errors, parameter estimates, and computation times among sequential least squares, linear approximation, and trust region interior point

Name	SLSQP	COBYLA	Trust region
Mean squared error	3538.92×10^7	108.27×10^7	25.69×10^7
β	0.485	0.716	0.675
σ	0.334	1.492	0.490
γ	0.070	0.083	0.070
ρ	0.0001	-0.004	0.002
τ	0.600	0.196	0.600
τ'	0.700	0.630	0.700
Time (sec)	151	813	2,720

1999). When the steps taken in the previous iteration are accurate in minimizing the objective function, a greater trust-region radius can be chosen to allow for bigger steps, which can potentially progress the optimization process towards the optimal solution faster, that is the optimization converges faster. By adjusting the trust-region radius at each iteration, the trust-region interior-point method aims to balance the trade-off between exploration and exploitation, finding a good compromise between taking large steps for faster convergence and ensuring accuracy in the optimization process.

Experiment setting

Initial starting points were selected randomly within a range of parameter values based on CDC reports (CDC, 2022). Details of the experiment setting were descried in Table 5.5. In the trust region method, the incubation and infectious periods are set to be within [2, 14] days and [5, 14] days, respectively (CDC, 2022). The initial value for r is 1 and the stopping criteria is 0.1×10^{-7} in the trust-region interior point method.

Table 5.5. Experiment setup for the trust-region interior-point method

Name	Range/Initial value
Iterations (k)	100
Initial value (r)	1
Stopping criteria per each iteration	$r < 0.1 \times 10^{-7}$
Total population (N)	19,378,102
Incubation period (days)	[2, 14]
Infectious period (days)	[5, 14]
Transmission rate (β)	[0,∞)
Exposed to infectious rate (σ)	[0,1]
Recovery rate (γ)	[0.07, 0.2]
Mortality rate (ρ)	[0.0001,0.1]
NPI effect 1 (τ)	[0,0.95]
NPI effect 2 (τ')	[0,0.95]

5.3.3 Alternating minimization

In the preliminary study, we investigated that the selected mortality rate by trust region interior point could be improved after investigating the graphical result. Therefore, we formulated an alternating minimization to target only the mortality rate while keeping the optimal values of the other parameters constant. Equation 5.4 includes the mean squared error between the reported- and estimated death counts.

$$\begin{aligned} \arg\min_{\rho} \quad & \frac{1}{T}\sum_{t=1}^{T} \|y_D(t) - h_\rho(t)\|^2 \\ \text{s.t.} \quad & \rho < 0.1, \\ & -\rho < 0, \end{aligned} \tag{5.4}$$

where $y_D(t)$ is the number of deceased population reported by CDC, $h_\rho(t)$ is the estimates of deceased cases, and ρ is a mortality rate. The upper bound of ρ is 0.1 (Ahmad et al.,

2021; Nabi, 2020). We applied the trust region interior point method to solve Equation 5.4.

5.3.4 Basic reproduction number of COVID-19

The reproduction number was estimated using a formula in Equation 5.5 (Van den Driessche, 2017; Jones, 2007).

$$R_0 = \frac{\sigma\beta}{\sigma(\gamma+\rho)} = \frac{\beta}{\gamma+\rho} \tag{5.5}$$

5.3.5 Model validation - bootstrap re-sampling

Using the parameter estimate obtained through trust region, we employed bootstrapping to create a confidence interval for the number of infected cases and the number of deaths. This was done to show the precision of the parameter estimates.

5.4 Result

5.4.1 Parameter estimates

We compared the parameter estimates with the results in Chowell et al. (2003). Because coronavirus is one of SARS virus and Chowell et al. (2003) studied a model of general SARS dynamics, the parameter estimates in their study can be used as a baseline in modeling COVID-19 (Carcione et al., 2020). Others studied the COVID-19 dynamics in New York City with consideration of vaccination, but the exact values of parameters were not found (Demongeot et al., 2022).

The initial transmission rates estimated in this study were shown to be similar with the ranges of baseline parameter values in published studies (Chowell et al., 2003; Carcione et al., 2020). The initial transmission rate was estimated by 0.73 by grid search, while the transmission rate was estimated by 0.675 by trust region interior point, com-

pared with 0.75 in Chowell et al. (2003) when $R_0 > 1$ without mitigation strategies. The incubation period was estimated by 3.8 days by grid search, while the incubation period was estimated by 2.04 days by trust region interior point, compared with 3 days in Chowell et al. (2003). In addition, the infectious period was estimated by 13.42 days by grid search, while the infectious period was estimated by 13.48 days by trust region, compared with 8 days in the baseline values. The mortality rate was estimated by 0.1×10^{-3} by grid search, while the mortality rate was estimated by 2.22×10^{-3} by trust region, compared to 1.44×10^{-3} in the baseline parameter.

Most importantly, the effect of school and business closures on March 23 was estimated by 77% by grid search, while the effect of school and business closures on March 23 was estimated by 40% by trust region interior point. The effect of metro closure was estimated by 14% by grid search, while the effect of metro closure was estimated by 30%. There was no literature found which quantified the effect of the New York on pause and the metro shutdown as of 2023. The effect of these policies were obtained by $100(1 - \tau's)$.

The average basic reproduction number R_0 obtained through grid search was 5.89, whereas the average R_0 calculated using the trust region interior point method was 5.85, as compared to the baseline study's value of 5.72.

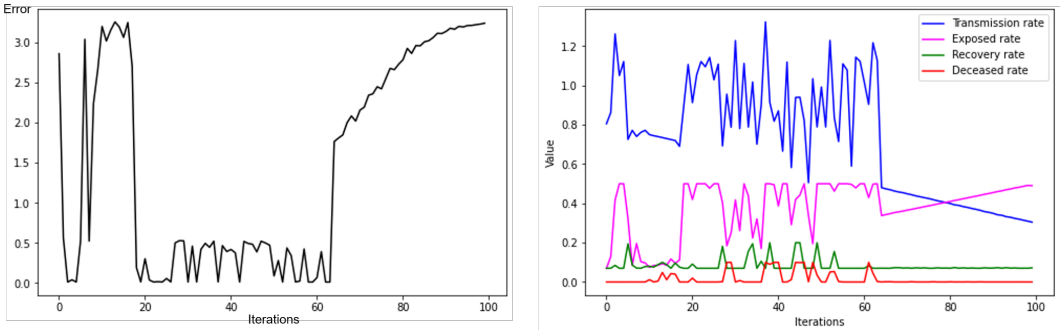
Table 5.6 summarizes the results from the grid search and trust region interior point method with comparison with the results in (Chowell et al., 2003). We recommend to consider the parameter estimates and the quantified effect of NPIs obtained through trust region interior point, since it exhibited smaller errors than the grid search.

In Figure 5.4, the errors and parameter estimates were illustrated according to 100 different initial values utilized in the trust region interior point method. The minimum was obtained when the initial values for parameters was $\theta = [\beta, \sigma, \gamma, \rho, \tau, \tau']^T = [0.675, 0.162, 0.07, 0.0001, 0.6, 0.7]^T$. Then, ρ, a mortality rate, was determined by solving a subsequent alternating minimization of the mean squared errors of the death count

with the initial value of 0.0016. We illustrate in Figure 5.5 (a) the beta distribution obtained from trust region with different starting points (b) the prediction of number of population in the state of SEIRD over a four hundred-day period, utilizing the optimal parameter that minimizes the mean squared error.

Table 5.6. Parameter estimates by grid search and trust region in this study compared to the values from a baseline study (Chowell et al. (2003))

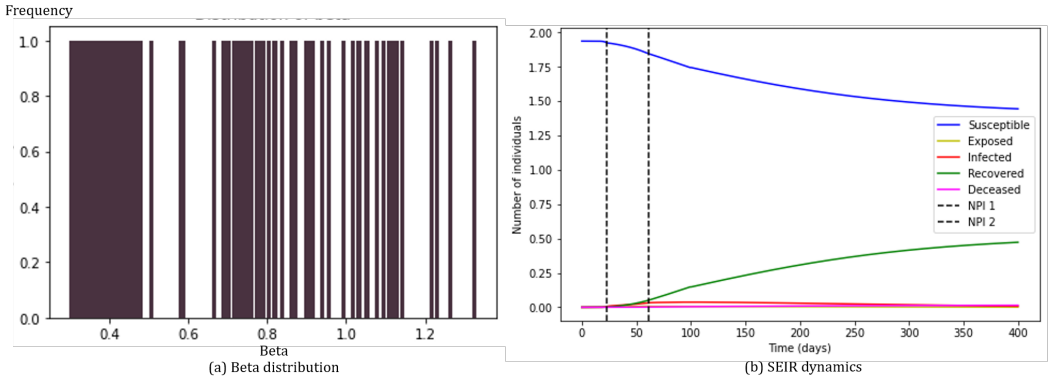
Name	**Grid Search**	**Trust Region**	**Baseline**
N	19,378,102	19,378,102	10,000,000
$\beta(t=0)$	0.730	0.675	0.750
$\beta(0 \leq t < 23)$	0.730	0.675	-
$\beta(23 \leq t < 61)$	0.562	0.405	-
$\beta(\mathrm{t} \geq 61)$	0.079	0.284	-
σ	0.2600 (=1/3.80)	0.4900 (=1/2.04)	0.3300 (=1/3)
γ	0.0700 ($\approx$1/13.42)*	0.0701 ($\approx$1/13.48)*	0.1250 (=1/8)
ρ	0.10×10^{-3}	2.22×10^{-3}	1.44×10^{-3}
τ	0.77	0.60	-
τ'	0.14	0.70	-
Error (I)	$1,080.00 \times 10^7$	25.69×10^7	-
Error (D)	369.29×10^7	3.36×10^7	-
Average R_0	5.89	5.85	5.72
Time (sec)	9,926	2,351	-



(a) Errors with different starting points

(b) Parameter values with different starting points

Figure 5.4. Trust region results - (a) the consequent mean squares of errors (b) parameter estimates with different initial starting points



(a) Beta distribution

(b) SEIR dynamics

Figure 5.5. Trust region and alternating optimization results – (a) Beta distribution and (b) prediction of coronavirus dynamics using SEIRD model with the optimal parameters from trust region and alternating optimization

5.4.2 Model validation

Bootstrap resampling

Based on the selected parameter values by trust region interior point, we constructed a confidence interval for the number of infected cases and the number of deceased cases. The result shows that the model by trust region is within 95% of confidence interval for 1,000 samples. Specifically, the confidence interval for the number of infected has a

lower bound of 167342.43 and a upper bound of 307133.06, which includes the median of the number of infected cases. Similarly, the median of the number of deceased cases was included in the 95% of confidence interval (i.e., lower bound = 4735.93, upper bound = 13899.89).

5.5 Discussion

5.5.1 Parameter estimates

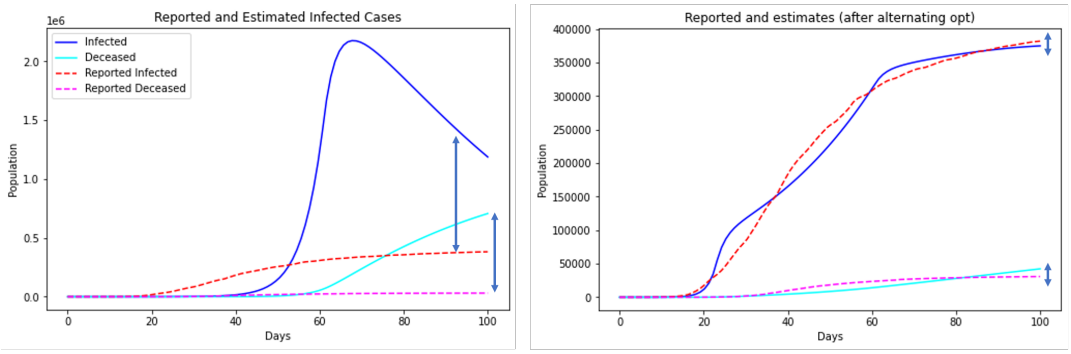
After comparing the MSE from grid search and trust region methods, the parameter estimates from trust region interior point had lower error than grid search. The parameter estimates by trust region interior point were compared to the baseline parameters of general SARS viruses in Chowell and Nishiura (2014). The initial transmission rate of 0.675 in this study is about 90% of the baseline transmission rate of 0.75. The latent period of 2.04 days is 0.86 days shorter than the baseline study of 3 days, while the infectious period of 13.48 days in this study is 5.48 days longer than the baseline study of 8 days. Further, the mortality rate of 0.0022 in this study is 0.78×10^{-3} higher than the baseline mortality rate of 0.0014. Intuitively, this investigation showed although the transmission rate of coronavirus is 10% less than the transmission rate of general SARS viruses, the infectious period of coronavirus is longer than general SARS viruses. This describes that COVID-19 symptoms last 5.48 days longer than usual SARS symptoms, possibly increasing the risk of transmission, causing the prolonged timeline of the COVID-19 outbreak. Notably, the mortality rate of coronavirus was 54% higher than general SARS viruses, showing the fatality of coronavirus. The average basic reproduction number R_0 was estimated as 5.85, which is about 2.72% larger than the baseline reproduction number.

The effectiveness of New York on Pause reduced the transmission rate by 40%, while the metro closure reduced the transmission rate by 30%. As of 2023, there have been

no studies that quantified the effectiveness of the New York on Pause and the metro closure, causing difficulty in direct comparison.

5.5.2 Comparison between grid search and trust region

The results indicated that the model chosen through trust region outperformed the model selected through grid search in terms of fitting empirical data with less error (as demonstrated in Figure 5.6) and achieving faster computation time.



(a) Comparison between the reported cases and the estimates by grid search

(b) Comparison between the reported cases and the estimates by trust region

Figure 5.6. Grid search and Trust region

5.5.3 Performance of grid search

Grid search explored 3,240,000 combinations of parameter values, requiring a substantial amount of computation time. In a preliminary study, we explored parameter values with a larger step size of 0.1 to approximate the range of best values. As a result, we determined that the transmission rate was roughly greater than 0.2. Therefore, we set the initial value for β to 0.2 instead of 0. After investigating the grid search results, we found that the best parameter values were similar to those from the trust region method.

5.5.4 Nonlinear constrained optimization to unconstrained optimization in the trust region interior point method

We have formulated a mathematical problem to optimize the model parameters for the SEIRD model for COVID-19. Theoretically, the constraints set in our optimization problem could be removed by incorporating a penalty function into the objective function. The primary motivation was to apply the existing solvers designed for unconstrained nonlinear optimization. To assess this, we conducted experiments using solvers available in the Python library for nonlinear optimization problems, specifically we used the optimize module within the Scipy library.

To convert the original problem's constraints into an unconstrained problem, a penalty function (such as $-\mu(\log\theta + \log(1-\theta))$) was introduced and added to the initial objective function. The trust region method was subsequently employed to determine the optimal values for various parameters, including transmission rate, exposure rate, recovery rate, fatality rate, the impact of the first non-pharmaceutical intervention in New York (known as New York on pause), and the impact of the second NPI, which involved the closure of the metro system.

5.5.5 Trust region and initial starting point

The selection of initial starting points had a significant impact on the optimal solution achieved through the trust region interior point method. To address this issue, we adopted a heuristic approach, trial and error method, to find the most suitable initial starting point that minimizes the objective value. Initially, we used an educated guess based on the outcome of grid search, and subsequently fine-tuned the initial values by making small adjustments within the constraints of each parameter. This approach aimed to narrow down the range of values and enhance the optimization process. In order to thoroughly explore the parameter space, we considered a hundred different

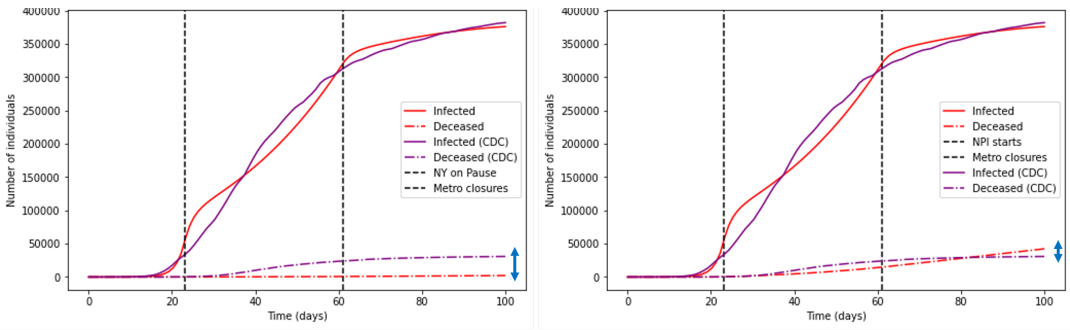
initial points and assessed their corresponding objective values. Since all six parameters were optimized simultaneously, their interdependence played a crucial role. Even if just one out of the six parameters deviated from the promising initial point, the trust region method could potentially become trapped in a local minimum. Therefore, it was essential to investigate a diverse set of initial starting points in the optimization process, utilizing the trust region method. Since we explored a hundred different combinations of initial values for the parameter vector, it allowed a comprehensive exploration of the parameter space.

5.5.6 Aggregated effect of recovered rate and deceased rate in the SEIRD

After examining the coronavirus dynamics using the optimal solution obtained from the primal optimization problem, it was investigated that the recovered- and deceased rate had an aggregated effect on the number of infected population in the objective function of mean squared error. Thus, the impact between the recovered rate and death rate on the loss function was not distinguishable due to the combined effect on the loss function. It was investigated that the mortality rate determined through the trust region method did not accurately align with the number of deaths reported by the CDC. Consequently, an additional problem of alternating minimization was formulated to identify the optimal death rate that minimized the mean squared errors between the reported and estimated number of deaths. As a result, we successfully developed an optimal model that fit empirical data, incorporating both the number of infected individuals and the number of deaths, while minimizing errors to the greatest extent possible.

5.5.7 Subsequent alternating minimization of the mortality rate

After optimizing all parameter values, the mortality rate was further calibrated though solving an alternating minimization problem with the objective function of the mean squared error between the reported number of deceased cases and the predicted number of deceased cases. By incorporating this additional calibration process, the modeling accuracy was enhanced as it effectively minimized errors not only in the counts of infected cases but also in the counts of deaths. Figure 5.7 presents a comparison between the outcomes obtained without using alternating minimization and those achieved with alternating minimization. The improvement in the error of death counts is visually represented by the blue arrow.



(a) Before alternating optimization – the reported cases (purple) and the estimates (red)

(b) After alternating optimization – the reported cases (purple) and the estimates (red)

Figure 5.7. Reduced error in the deceased cases – (a) Before alternating optimization (b) after alternating optimization

5.5.8 Measure of the effects of nonpharmaceutical interventions

Transmission rate was modeled as a dynamic variable ($\beta(t)$) affected by public policies (τ) such that $\beta(t) = \tau\beta(t-1)$ when a public policy is implemented at time t. This idea allowed to fit empirical data in a realistic scenario and enables to measure the public policies effect. We updated the variable twice, on March 23 and April 30,

after contemplating the school closures and metro closures ordered in New York on each respective date. Deployment of additional public policy decreased transmission rate further, resulting in slowing down the transmission chain after 61-70 days. This finding aligns with the official announcement of the first decline in the hospitalization rate on May 20^{th}.

6. CONCLUSION

The purpose of this study was to integrate social media data into epidemic modeling to improve prediction accuracy. The study analyzed two sets of empirical data collected from Twitter and the Centers for Disease Control and Prevention, covering the period from March 2020 to May 2020. The first research question attempted to answer how to evaluate the extent of individual adoption of preventive behaviors during an epidemic. Text analytics was employed in analyzing Twitter data. Furthermore, a self-training machine learning model, consisting of support vector machines and k-nearest neighbor, was developed to classify individual use of prevention behaviors based on the social media data. The second research question addressed improvement of can we improve the accuracy of epidemic modeling by incorporating the effect of preventive behaviors on transmission. The classical compartmental differential equations system was applied to model COVID-19 dynamics. For parameter estimation, both heuristic algorithms and nonlinear optimization techniques were employed, chosen based on the minimum mean squared error. Specifically, grid search and the trust-region interior point method were applied in parameter estimation. These two analyses help to address our research questions.

6.1 Evaluation of the Compliance with Individual Prevention Behaviors

The novel idea of predicting individual preventive behaviors was investigated. Predicting pre-defined preventive behaviors, in the social media data was considered as a form of semi-supervised learning. In addressing the proposed problem, this study developed a self-training machine learning algorithm, called `PRecomm`, which stands for prediction and recommendation.

Empirical data, in the form of tweets, was collected from Twitter aligned with the timeline during which the COVID-19 dynamics were analyzed (from March 1, 2020, to

May 31, 2020). Techniques in text analytics, such as regular expressions and TF-IDF vectorization, were used for analyzing the text. Additionally, the study addressed the challenge of imbalance within the collected data by removing excessive tweets related to each preventive behavior and by collecting more tweets in under-represented classes.

Findings in this study indicated that the performance of self-training machine learning was affected by the selection of the classifier. It was shown that self-trained PRecomm processed 4,764 tweets with 0.97 F-1 score in less than 1,827 seconds. The study revealed that the support vector machine outperforms the random forest in terms of prediction accuracy, particularly the F-1 score, as well as computation time. Additionally, the study found that k-nearest neighbor was useful in predicting multiple preventive behaviors due to its fast computation time. The study suggested that the results from both the support vector machine and k-nearest neighbor could be utilized together in the real world for a flexible and resilient response to emergency situations.

The research also emphasized the challenge of imbalance in the collected data. The analytical results of this study indicated that even a small imbalance in the collected data led to convergence towards a majority label after self-training machine learning. Therefore, the research emphasized that addressing such imbalance should take precedence over self-training.

Another finding indicated that self-training machine learning did not improve the F-1 score compared to classical machine learning. However, it did improve both average prediction probabilities in support vector machines and random forests, as well as the average similarity scores in k-nearest neighbor models. Furthermore, the performance of classical machine learning was as high as the performance of self-training machine learning, partially due to well-encoded and well-balanced data. However, encoding and balancing training data required meticulous work. Therefore, it was recommended to experiment with how self-training performs without such sedulous work on training data to discover the true benefits of self-training.

Based on the analytical results, approximately 25-30% of individuals were likely to follow the preventive measures of hygiene, travel bans, and quarantine, whereas only 5-13% of the population was likely to follow mask-wearing and distancing. Quantitative results from this study offer insights for improving deployment of prevention strategies with low compliance. For example, policy makers can develop strategies (e.g., deployment of personal protective tools, such as masks) to help improve prevention strategies with weak compliance. Furthermore, disaster relief organizations can provide personalized social media contents to influence individuals with low compliance.

6.2 Incorporation of the Effect of Preventive Behaviors with Epidemic Modeling

This study adapted a SEIRD model to examine the dynamics of the coronavirus. The model investigation accounted for the influence of specific public health policies, including closures of schools, stores, and subways. The transmission rate in this study was characterized by a dynamic variable, allowing for the estimation of the effects of non-pharmaceutical interventions (NPIs), considering the temporal implementation of public policies in New York. To estimate the model parameters, the study employed grid search, a heuristic method and nonlinear optimization using trust region interior point method and alternating optimization. The study recommended prioritizing the results obtained from the trust region method rather than the grid search due to its lower error. For model validation, bootstrap resampling was utilized.

For New York between March and May of 2020, average reproduction number R_0 was estimated as [5.85,5.89]. This average basic reproduction number was approximately 2.72% higher than the baseline reproduction number (R_0=0.72) of the general SARS virus studied in Chowell et al. (2003). This study estimated effect of the New York lockdown at 40%, indicating that public policy reduced the transmission rate by 40%. Meanwhile, the effect of metro closure was estimated at 30%, suggesting that metro

closure reduced the transmission rate by 30%. The analytical results suggest that the lockdown had a greater effect than the metro closure in reducing COVID-19 transmission rate.

Based on the results from both the grid search and the trust region method, the best-fitting model was obtained with an initial transmission rate of [0.675,0.730], an exposed rate of [0.26,0.49], and a recovered rate of 0.07, resulting in an incubation period of approximately 2.04 to 3.80 days and an infectious period of approximately 13.42 to 13.48 days. This result estimates the infectious period of COVID-19 to be 5 days longer than that of the typical SARS virus, possibly prolonging the duration of the pandemic. Furthermore, the mortality rate of COVID-19 was also about 54% higher than the mortality rate of the typical SARS viruses, consistent with the high fatality of coronavirus.

In conclusion, the study emphasized the importance of incorporating empirical data regarding public policies, such as the types of public policies, precise implementation dates, and localized considerations, into the modeling and analysis procedures for epidemic dynamics. This integration can enhance the prediction accuracy by incorporating realistic scenarios. In parameter estimation, the single objective function can be improved by solving an additional optimization problem or adding a multi-objective function to properly estimate the deceased rate. It is also critical to consider a wide range of initial starting points when applying trust-region method. The study's findings demonstrate the potential to be applied to diverse regions and different virus variants, thereby offering important support to policymakers engaged in disaster management.

6.3 Contribution

This research demonstrated the novel use of social media data to predict individual compliance with prevention behaviors and integrating the effect of prevention behaviors into epidemic modeling. This research provides an important foundation for integrating

contextual data into epidemic modeling. Classical epidemic modeling only considers the magnitude of population size in each compartment. This is based on the assumption of homogeneity of the population, and can be relaxed by considering a heterogeneous population with different statuses and conditions, as learned from contextual information. In application, contextual information can provide real-time awareness of individual behavior and health status during outbreaks of diseases. As seen in this research, the implementation level of non-pharmaceutical interventions can be learned by the individual compliance level through social media data. This result can be used in epidemic modeling to consider the effect of non-pharmaceutical intervention by replacing the predicted population-level compliance with prevention behaviors learned from contextual data, then weighted to calibrate the model accuracy. This idea is novel and this research provides an important cornerstone for the feasibility of integrating contextual data into epidemic modeling for practical and resilient real-time responses to disasters.

7. LIMITATION & FUTURE WORK

Evaluation of the Individual Preventive Behaviors

The research should be improved in three aspects: (1) incorporating qualitative data, (2) broadening the data, (3) and methodology. Most importantly, the study needs to be combined with qualitative data such as interviews, observations, or survey data to verify the prediction of individual prevention behaviors. Secondly, social media data should be expanded to include data from multiple platforms while also increasing the data size to reduce selection bias and improve generalizability. The labels for non-prevention should be considered in encoding process. Multiple labels should be incorporated as well. Thirdly, term frequency–inverse document frequency vectorizer should be compared to other methods, such as BERT transformer, for representation of context. Fourthly, feature selection should also be conducted in improving model performance. Finally, it would be worth investigating alter native labels during encoding, for example labeling non-compliance and use of multiple prevention behaviors.

Incorporation of the Effect of Preventive Behaviors with Epidemic Modeling

The research should examine enhancements to improve the prediction accuracy. First, this research should incorporate additional time points, such as weekly updates, to enable more frequent updates of the transmission rate. To accomplish this, the integration of real-time communication systems within disaster management and epidemic modeling systems is crucial to enable more timely and frequent incorporation of real-world data. Furthermore, the research should take into account the emergence of new coronavirus variants over time. While the study made attempts to predict the long-term dynamics of COVID-19 using empirical data from the initial three months, it may not accurately reflect disease progression beyond the time when a new coronavirus variant emerged. Lastly, it is recommended that future research focuses on examining the ef-

fects of local interventions in each affected area, while considering geo-spatial limitations. This would enable a comprehensive understanding of the impact of interventions within specific regions.

Heterogeneous Epidemic Modeling

The challenges of incorporating social media data into the SEIRD model should also be addressed in future work. The first step towars this goal is to introduce heterogeneous populations, those who use social media denoted as S_1, and those who do not use social media as S_0. The transmission rates from each susceptible population to the exposed compartment can be modeled by β_0 and β_1 for S_0 and S_1, respectively. The transmission rates are then influenced by the effects of preventive behaviors, denoted by τ and τ' for β_0 and β_1. Figure 7.1 illustrates the proposed heterogeneous SEIRD model.

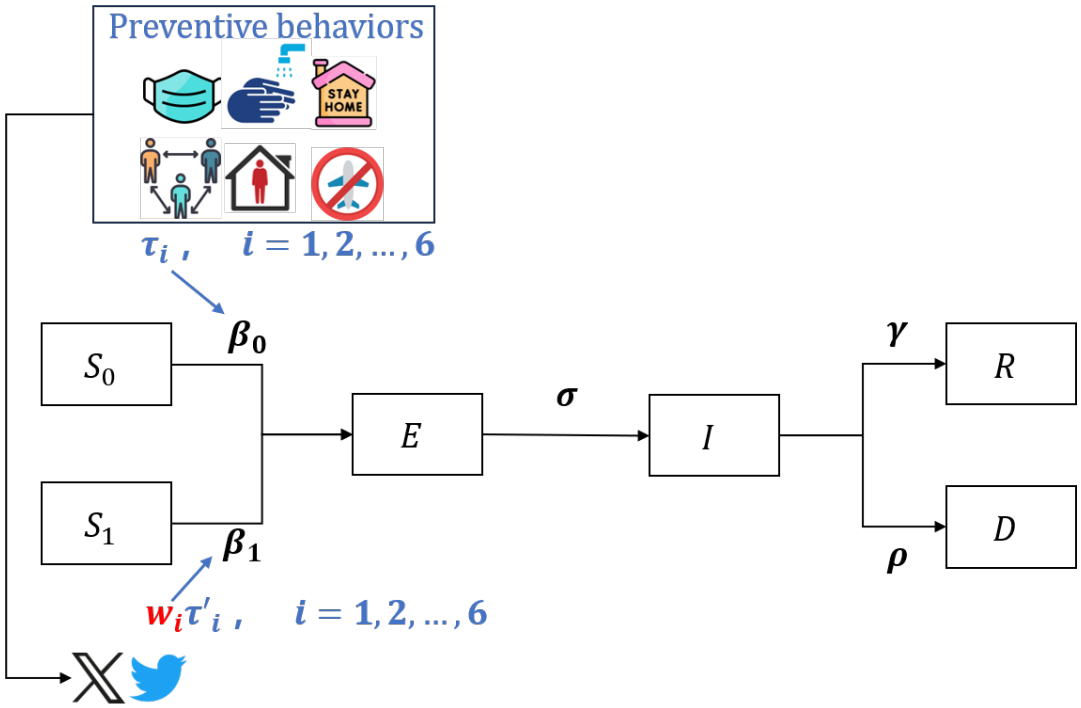


Figure 7.1. Heterogeneous SEIRD with variables and weights

Bibliography

Abou-Ismail, A., 2020: Compartmental models of the covid-19 pandemic for physicians and physician-scientists. *SN comprehensive clinical medicine*, **2 (7)**, 852–858.

Ahmad, F. B., J. A. Cisewski, A. Miniño, and R. N. Anderson, 2021: Provisional mortality data—United States, 2020. *Morbidity and Mortality Weekly Report*, **70 (14)**, 519.

Antonelli, M., J. C. Pujol, T. D. Spector, S. Ourselin, and C. J. Steves, 2022: Risk of long COVID associated with delta versus omicron variants of SARS-CoV-2. *The Lancet*, **399 (10343)**, 2263–2264.

Bae, S., E. Sung, and O. Kwon, 2021: Accounting for social media effects to improve the accuracy of infection models: Combatting the COVID-19 pandemic and infodemic. *European Journal of Information Systems*, **30 (3)**, 342–355.

Baevski, A., and A. Mohamed, 2020: Effectiveness of self-supervised pre-training for ASR. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 7694–7698.

Baeza-Yates, R., 1999: Modern Information Retrieval. *Addison Wesley Google Schola*, **2**, 127–136.

Bamman, D., and N. Smith, 2015: Contextualized Sarcasm Detection on Twitter. *Proceedings of the International AAAI conference on Web and Social Media*, Vol. 9, 574–577.

Bandura, A., 1984: *Representing personal determinants in causal structures.* American Psychological Association.

BBC, 2016: Artificial intelligence: Go master Lee Se-dol wins against AlphaGo program. BBC News Online, URL https://www.bbc.co.uk/news/technology-35797102.

Beckley, R., C. Weatherspoon, M. Alexander, M. Chandler, A. Johnson, and G. S. Bhatt, 2013: Modeling epidemics with differential equations. *Tennessee State University Internal Report*.

Berthelot, D., N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, 2019: Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, **32**.

Bertsimas, D., and J. N. Tsitsiklis, 1997: *Introduction to linear optimization*, Vol. 6. Athena Scientific Belmont, MA.

Bharti, S. K., K. S. Babu, and S. K. Jena, 2015: Parsing-based sarcasm sentiment recognition in Twitter data. *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, 1373–1380.

Bird, S., E. Klein, and E. Loper, 2009: *Natural language processing with python.* O'Reilly Media.

Bishop, C. M., and N. M. Nasrabadi, 2006: *Pattern recognition and machine learning*, Vol. 4. Springer.

Blei, D. M., A. Y. Ng, and M. I. Jordan, 2003: Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3 (Jan)**, 993–1022.

Boyd, S. P., and L. Vandenberghe, 2004: *Convex Optimization*. Cambridge University Press.

Brauer, F., P. Van den Driessche, J. Wu, and L. J. Allen, 2008: *Mathematical epidemiology*, Vol. 1945. Springer.

Burki, T. K., 2021: The race between vaccination and evolution of COVID-19 variants. *The Lancet Respiratory Medicine*, **9 (11)**, e109.

Burton, S. H., K. W. Tanner, C. G. Giraud-Carrier, J. H. West, and M. D. Barnes, 2012: "Right time, right place" health communication on Twitter: Value and accuracy of location information. *Journal of Medical Internet Research*, **14 (6)**, e2121.

Cambria, E., B. Schuller, Y. Xia, and C. Havasi, 2013: New avenues in opinion mining and sentiment analysis. *IEEE Intelligent systems*, **28 (2)**, 15–21.

Carcione, J. M., J. E. Santos, C. Bagaini, and J. Ba, 2020: A simulation of a COVID-19 epidemic based on a deterministic SEIR model. *Frontiers in Public Health*, **8**, 230.

Cavallaro, C., A. Bujari, L. Foschini, G. Di Modica, and P. Bellavista, 2021: Measuring the impact of COVID-19 restrictions on mobility: A real case study from Italy. *Journal of Communications and Networks*, **23 (5)**, 340–349.

CDC, 2019: COVID-19 Data Tracker. CDC, URL https://covid.cdc.gov/covid-data-tracker/#datatracker-home.

CDC, 2020: How to Protect Yourself & Others. CDC, URL https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html.

CDC, 2020: United States COVID-19 Cases and Deaths by State over Time. Centers for Disease Control and Prevention, [Online; accessed 20-October-2020], https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-over-Time.

CDC, 2021: How to protect yourself and others. Centers for Disease Control and Prevention, URL https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html.

CDC, 2021: United States COVID-19 Cases and Deaths by State over Time. CDC, URL https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36/data.

CDC, 2022: Ending Isolation and Precautions for People with COVID-19: Interim Guidance. Centers for Disease Control and Prevention, https://www.cdc.gov/coronavirus/2019-ncov/hcp/duration-isolation.html Accessed 08.31.2022.

CDC, 2023a: COVID-19 Timeline. CDC, URL https://www.cdc.gov/museum/timeline/covid19.html#print.

CDC, 2023b: End of the Federal COVID-19 Public Health Emergency (PHE) Declaration. Centers for Disease Control and Prevention, URL https://archive.cdc.gov/#/details?url=https://www.cdc.gov/coronavirus/2019-ncov/your-health/end-of-phe.html.

Chan, D. K., C.-Q. Zhang, and K. Weman-Josefsson, 2021: Why people failed to adhere to COVID-19 preventive behaviors? Perspectives from an integrated behavior change model. *Infection Control & Hospital Epidemiology*, **42 (3)**, 375–376.

Chatfield, A. T., and C. G. Reddick, 2018: All hands on deck to tweet# sandy: Networked governance of citizen coproduction in turbulent times. *Government Information quarterly*, **35 (2)**, 259–272.

Chen, E., K. Lerman, and E. Ferrara, 2020a: Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set. *JMIR Public Health and Surveillance*, **6 (2)**, e19 273.

Chen, Q., C. Min, W. Zhang, G. Wang, X. Ma, and R. Evans, 2020b: Unpacking the black box: How to promote citizen engagement through government social media during the covid-19 crisis. *Computers in human behavior*, **110**, 106 380.

Chinazzi, M., and Coauthors, 2020: The effect of travel restrictions on the spread of the 2019 novel coronavirus COVID-19 outbreak. *Science*, **368 (6489)**, 395–400.

Cho, H., 2016: The casualty transportation of Ebola outbreak in Liberia: A simulation study. Master thesis, Purdue University, URL https://docs.lib.purdue.edu/open_access_theses/760/.

Cho, H., and R. Shehab, 2023: Individual engagement in prevention behaviors during COVID-19: Self-training machine learning using Twitter data. *Proceedings of the International AAAI conference on AI for Behavior Change*.

Chowell, G., P. W. Fenimore, M. A. Castillo-Garsow, and C. Castillo-Chavez, 2003: SARS outbreaks in Ontario, Hong Kong and Singapore: The role of diagnosis and isolation as a control mechanism. *Journal of Theoretical Biology*, **224 (1)**, 1–8.

Chowell, G., N. W. Hengartner, C. Castillo-Chavez, P. W. Fenimore, and J. M. Hyman, 2004: The basic reproductive number of Ebola and the effects of public health measures: The cases of Congo and Uganda. *Journal of Theoretical Biology*, **229 (1)**, 119–126.

Chowell, G., and H. Nishiura, 2014: Transmission dynamics and control of Ebola virus disease (EVD): A review. *BMC Medicine*, **12**, 1–17.

Christopher Manning, 2020: Artificial Intelligence Definitions. Standford University, URL https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf.

Cinelli, M., and Coauthors, 2020: The COVID-19 social media infodemic. *Scientific reports*, **10 (1)**, 1–10.

Clement, J., 2020: Number of global social network users 2017–2025. *Retrieved from the Statista website: https://www. statista. com/statistics/278414/number-of-worldwide-social-network-users*.

Cover, T., and P. Hart, 1967: Nearest neighbor pattern classification. *IEEE transactions on information theory*, **13 (1)**, 21–27.

Dasarathy, B. V., 1991: Nearest neighbor (NN) norms: NN pattern classification techniques. *IEEE Computer Society Tutorial*.

Davidov, D., and A. Rappoport, 2006: Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 297–304.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, 1990: Indexing by latent semantic analysis. *Journal of the American society for information science*, **41 (6)**, 391–407.

Demongeot, J., Q. Griette, P. Magal, and G. Webb, 2022: Modeling vaccine efficacy for COVID-19 outbreak in New York City. *Biology*, **11 (3)**, 345.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova, 2018: Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Diaz, P., P. Constantine, K. Kalmbach, E. Jones, and S. Pankavich, 2018: A modified SEIR model for the spread of Ebola in Western Africa and metrics for resource allocation. *Applied Mathematics and Computation*, **324**, 141–155.

Diekmann, O., J. A. P. Heesterbeek, and J. A. Metz, 1990: On the definition and the computation of the basic reproduction ratio r 0 in models for infectious diseases in heterogeneous populations. *Journal of mathematical biology*, **28 (4)**, 365–382.

Donnelly, C. A., M. R. Malik, A. Elkholy, S. Cauchemez, and M. D. Van Kerkhove, 2019: Worldwide reduction in MERS cases and deaths since 2016. *Emerging Infectious Diseases*, **25 (9)**, 1758.

Enns, E. A., and Coauthors, 2020: Modeling the impact of social distancing measures on the spread of sars-cov-2 in minnesota. Tech. rep., University of Minnesota, Minnesota, 16 pp.

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

Ferguson, T. S., 1973: A Bayesian analysis of some nonparametric problems. *The annals of statistics*, 209–230.

Foulds, J., and P. Smyth, 2011: Multi-instance mixture models and semi-supervised learning. *Proceedings of the 2011 SIAM International Conference on Data Mining*, SIAM, 606–617.

Friji, H., R. Hamadi, H. Ghazzai, H. Besbes, and Y. Massoud, 2021: A generalized mechanistic model for assessing and forecasting the spread of the COVID-19 pandemic. *IEEE Access*, **9**, 13 266–13 285.

Ghosh, D., and R. Guha, 2013: What are we 'tweeting'about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science*, **40 (2)**, 90–102.

Griffiths, T. L., and M. Steyvers, 2004: Finding scientific topics. *Proceedings of the National academy of Sciences*, **101 (suppl_1)**, 5228–5235.

Guy, G., and Coauthors, 2021: Association of State-Issued Mask Mandates and Allowing On-Premises Restaurant Dining with County-Level COVID-19 Case and Death

Growth Rates — United States, March 1–December 31, 2020. *MMWR. Morbidity and Mortality Weekly Report*, **70**, doi:10.15585/mmwr.mm7010e3.

Haug, N., and Coauthors, 2020: Ranking the effectiveness of worldwide COVID-19 government interventions. *Nature Human Behaviour*, **4 (12)**, 1303–1312.

He, Q., L. Wang, and B. Liu, 2007: Parameter estimation for chaotic systems by particle swarm optimization. *Chaos, Solitons & Fractals*, **34 (2)**, 654–661.

Hsu, C.-W., C.-C. Chang, C.-J. Lin, and Coauthors, 2003: A practical guide to support vector classification. Taipei.

Hu, T., J. Luo, and W. Liu, 2018: Life in the Matrix: Human Mobility Patterns in the Cyber Space. *Twelfth International AAAI Conference on Web and Social Media*.

Huang, C., and Coauthors, 2020: Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The lancet*, **395 (10223)**, 497–506.

Insider Intelligence, 2020: Coronavirus Impact: US Adults Who Have Significantly Increased* Their Usage of Social Media, by Platform. Insider Intelligence, (Accessed: 2021-10-30), http://www.insiderintelligence.com/chart/236819.

Ivanov, D., C. S. Tang, A. Dolgui, D. Battini, and A. Das, 2021: Researchers' perspectives on Industry 4.0: Multi-disciplinary analysis and opportunities for operations management. *International Journal of Production Research*, **59 (7)**, 2055–2078.

Jones, J. H., 2007: Notes on R0. *California: Department of Anthropological Sciences*, **323**, 1–19.

Kaner, J., and S. Schaack, 2016: Understanding Ebola: The 2014 epidemic. *Globalization and Health*, **12 (1)**, 1–7.

Katella, K., M. Vazquez, E. Shapiro, and T. Murray, 2023: Omicron, Delta, Alpha, and More: What To Know About the Coronavirus Variants. Yale Medicine, URL https://www.yalemedicine.org/news/covid-19-variants-of-concern-omicron.

Kermack, W. O., 1927: A contributions to the mathematical theory of epidemics. *Proc. R. Soc. Edinburgh A*, Vol. 115, 700.

Kermack, W. O., and A. G. McKendrick, 1927: A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, **115 (772)**, 700–721.

Korea Baduk Association, 2016: Saedol Lee vs AlphaGo, 'Google DeepMind Challenge Match' press meeting will be held in Four Seasons Hotel Seoul on the 9th next month, the curtain rises for the historical match. BBC News Online, URL https://web.archive.org/web/20160303210212/http://www.baduk.or.kr/news/report_view.asp?news_no=1671.

Kumar, S., C. Xu, N. Ghildayal, C. Chandra, and M. Yang, 2021: Social media effectiveness as a humanitarian response to mitigate influenza epidemic and COVID-19 pandemic. *Annals of Operations Research*, 1–29.

Kursan Milaković, I., 2021: Purchase experience during the COVID-19 pandemic and social cognitive theory: The relevance of consumer vulnerability, resilience, and adaptability for purchase satisfaction and repurchase. *International Journal of Consumer Studies*, **45 (6)**, 1425–1442.

Kwon, K.-S., J.-I. Park, Y. J. Park, D.-M. Jung, K.-W. Ryu, and J.-H. Lee, 2020: Evidence of long-distance droplet transmission of SARS-CoV-2 by direct air flow in a restaurant in Korea. *Journal of Korean Medical Science*, **35 (46)**.

Lamsal, R., 2021: Design and analysis of a large-scale COVID-19 tweets dataset. *Applied Intelligence*, **51 (5)**, 2790–2804.

Latah, M., and L. Toker, 2020: An efficient flow-based multi-level hybrid intrusion detection system for software-defined networks. *CCF Transactions on Networking*, **3 (3-4)**, 261–271.

Lee, S., C. Hwang, and M. J. Moon, 2020: Policy learning and crisis policy-making: Quadruple-loop learning and COVID-19 responses in South Korea. *Policy and Society*, **39 (3)**, 363–381.

Li, X., Q. Liu, and Coauthors, 2020: Social media use, eHealth literacy, disease knowledge, and preventive behaviors in the COVID-19 pandemic: Cross-sectional study on Chinese netizens. *Journal of Medical Internet Research*, **22 (10)**, e19 684.

Lin, J., H. Aprahamian, and H. El-Amine, 2023: Optimal unlabeled set partitioning with application to risk-based quarantine policies. *IISE Transactions*, 1–13.

Liu, P. L., 2021: COVID-19 information on social media and preventive behaviors: Managing the pandemic through personal responsibility. *Social Science & Medicine*, **277**, 113 928.

Liu, Y., C. Morgenstern, J. Kelly, R. Lowe, and M. Jit, 2021: The impact of non-pharmaceutical interventions on SARS-CoV-2 transmission across 130 countries and territories. *BMC Medicine*, **19 (1)**, 1–12.

Loria, S., 2017: Advanced Usage: Overriding Models and the Blobber Class. Retrieved from TextBlob: https://textblob. readthedocs. io/en/dev , URL https://textblob.readthedocs.io/en/latest/advanced_usage.html#sentiment-analyzers.

McCarthy, J., M. L. Minsky, N. Rochester, and C. E. Shannon, 1955: A proposal for the dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Magazine*, **27 (4)**, 12–12.

McCulloch, W. S., and W. Pitts, 1943: A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, **5**, 115–133.

Meier, P., 2010: The unprecedented role of sms in disaster response: Learning from haiti. *SAIS Review of International Affairs*, **30 (2)**, 91–103.

Merchant, R. M., and N. Lurie, 2020: Social media and emergency preparedness in response to novel coronavirus. *Jama*, **323 (20)**, 2011–2012.

Microsoft Azure, 2024: Artificial intelligence (AI) vs. machine learning (ML) . Microsoft, URL https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/artificial-intelligence-vs-machine-learning#:~:text=Machine%20learning%20is%20an%20application,its%20own%2C%20based%20on%20experience.

Miller, G. A., 1995: WordNet: A lexical database for English. *Communications of the ACM*, **38 (11)**, 39–41.

Mirończuk, M. M., and J. Protasiewicz, 2018: A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, **106**, 36–54.

Moeini, B., S. Bashirian, A. R. Soltanian, A. Ghaleiha, and M. Taheri, 2019: Examining the effectiveness of a web-based intervention for depressive symptoms in female adolescents: Applying social cognitive theory. *Journal of Research in Health Sciences*, **19 (3)**, e00 454.

Murshed, B. A. H., J. Abawajy, S. Mallappa, M. A. N. Saif, S. M. Al-Ghuribi, and F. A. Ghanem, 2022: Enhancing Big Social Media Data Quality for Use in Short-Text Topic Modeling. *IEEE Access*, **10**, 105 328–105 351.

Nabi, K. N., 2020: Forecasting COVID-19 pandemic: A data-driven analysis. *Chaos, Solitons & Fractals*, **139**, 110 046.

NewYork, 2020: COVID-19 orders. New York, https://www.governor.ny.gov/news/governor-cuomo-signs-new-york-state-pause-executive-order, accessed 15.01.2022.

Nocedal, J., and S. J. Wright, 1999: *Numerical optimization*, page 66–69. Springer.

Nsoesie, E. O., R. J. Beckman, S. Shashaani, K. S. Nagaraj, and M. V. Marathe, 2013: A simulation optimization approach to epidemic forecasting. *PloS One*, **8 (6)**, e67 164.

Pang, B., L. Lee, and Coauthors, 2008: Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, **2 (1–2)**, 1–135.

Pham, Q.-V., D. C. Nguyen, T. Huynh-The, W.-J. Hwang, and P. N. Pathirana, 2020: Artificial intelligence (AI) and big data for coronavirus COVID-19 pandemic: A survey on the state-of-the-arts. *IEEE Access*, **8**, 130 820–130 839.

Qazi, U., M. Imran, and F. Ofli, 2020: GeoCoV19: A dataset of hundreds of millions of multilingual COVID-19 tweets with location information. *SIGSPATIAL Special*, **12 (1)**, 6–15.

Rahmani, R., and S. A. Goldman, 2006: Missl: Multiple-instance semi-supervised learning. *Proceedings of the 23rd international conference on Machine learning*, 705–712.

Raiffa, H., R. Schlaifer, and Coauthors, 1961: *Applied statistical decision theory*. Wiley New York.

Rosenblatt, F., and Coauthors, 1962: *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*, Vol. 55. Spartan books Washington, DC.

Ross, S. M., 2014: *Introduction to probability models*. Academic press.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams, 1986: Learning internal representations by error propagation, parallel distributed processing, explorations in the microstructure of cognition, ed. de rumelhart and j. mcclelland. vol. 1. 1986. *Biometrika*, **71**, 599–607.

Sarkar, D., 2016: *Text Analytics with python*. Springer.

Seo, M., A. Kembhavi, A. Farhadi, and H. Hajishirzi, 2016: Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Seth, N., 2021: Topic Modeling and Latent Dirichlet Allocation (LDA) using Gensim and Sklearn . Analytics Vidhya, URL https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/.

Sharf, Z., and S. U. Rahman, 2018: Performing natural language processing on roman urdu datasets. *International Journal of Computer Science and Network Security*, **18 (1)**, 141–148.

Shim, E., A. Tariq, W. Choi, Y. Lee, and G. Chowell, 2020: Transmission potential and severity of COVID-19 in South Korea. *International Journal of Infectious Diseases*, **93**, 339–344.

Shin, H.-Y., 2021: A multi-stage SEIR (D) model of the COVID-19 epidemic in Korea. *Annals of Medicine*, **53 (1)**, 1160–1170.

Steven, Loria , 2017: TextBlob: Simplified Text Processing. GitHub, URL https://github.com/sloria/TextBlob.

Sze, V., Y.-H. Chen, J. Emer, A. Suleiman, and Z. Zhang, 2017: Hardware for machine learning: Challenges and opportunities. *2017 IEEE custom integrated circuits conference (CICC)*, IEEE, 1–8.

Tang, B., X. Wang, Q. Li, N. L. Bragazzi, S. Tang, Y. Xiao, and J. Wu, 2020: Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions. *Journal of Clinical Medicine*, **9 (2)**, 462.

Teklu, S. W., and B. S. Kotola, 2023: A dynamical analysis and numerical simulation of COVID-19 and HIV/AIDS co-infection with intervention strategies. *Journal of Biological Dynamics*, **17 (1)**, 2175 920.

Tsay, C., F. Lejarza, M. A. Stadtherr, and M. Baldea, 2020: Modeling, state estimation, and optimal control for the US COVID-19 outbreak. *Scientific Reports*, **10 (1)**, 10 711.

Tsur, O., D. Davidov, and A. Rappoport, 2010: ICWSM—a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. *Fourth International AAAI Conference on Weblogs and Social Media*.

Twitter, 2019: Twitter API Documentation. Twitter, URL https://developer.twitter.com/en/docs/twitter-api.

United States Census, 2020: Population Estimates. United States Census Bureau, https://www.census.gov/quickfacts/NY Accessed 01.20.2021.

Valladares, L., V. Nino, K. Martínez, D. Sobek, D. Claudio, and S. Moyce, 2022: Optimizing patient flow, capacity, and performance of COVID-19 vaccination clinics. *IISE Transactions on Healthcare Systems Engineering*, **12 (4)**, 275–287.

Van den Driessche, P., 2017: Reproduction numbers of infectious disease models. *Infectious Disease Modelling*, **2 (3)**, 288–303.

Van Egeren, D., A. Novokhodko, M. Stoddard, U. Tran, B. Zetter, M. S. Rogers, D. Joseph-McCarthy, and A. Chakravarty, 2021: Controlling long-term SARS-CoV-2 infections can slow viral evolution and reduce the risk of treatment failure. *Scientific Reports*, **11 (1)**, 22 630.

Wächter, A., and L. T. Biegler, 2006: On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, **106**, 25–57.

WHO, 2021: WHO announces simple, easy-to-say labels for SARS-CoV-2 Variants of Interest and Concern. wHO, URL https://www.who.int/news/item/31-05-2021-who-announces-simple-easy-to-say-labels-for-sars-cov-2-variants-of/-interest-and-concern.

Widrow, B., M. E. Hoff, and Coauthors, 1960: Adaptive switching circuits. *IRE WESCON convention record*, New York, Vol. 4, 96–104.

Wikipedia, 2021: COVID-19 lockdowns. WIKIPEDIA, https://en.wikipedia.org/wiki/COVID-19_lockdowns.

Wilder-Smith, A., and D. O. Freedman, 2020: Isolation, quarantine, social distancing and community containment: Pivotal role for old-style public health measures in the novel coronavirus 2019-nCoV outbreak. *Journal of Travel Medicine*, **27 (2)**, taaa020.

Yang, H., S. Kumara, S. T. Bukkapatnam, and F. Tsung, 2019: The internet of things for smart manufacturing: A review. *IISE Transactions*, **51 (11)**, 1190–1216.

Yarsky, P., 2021: Using a genetic algorithm to fit parameters of a COVID-19 SEIR model for US states. *Mathematics and Computers in Simulation*, **185**, 687–695.

Zhou, X., W. Liang, Z. Luo, and Y. Pan, 2021: Periodic-aware intelligent prediction model for information diffusion in social networks. *IEEE Transactions on Network Science and Engineering*, **8 (2)**, 894–904.

Zhu, X. J., 2005: Semi-supervised learning literature survey.

Zoph, B., G. Ghiasi, T.-Y. Lin, Y. Cui, H. Liu, E. D. Cubuk, and Q. Le, 2020: Rethinking pre-training and self-training. *Advances in Neural Information Processing Systems*, **33**, 3833–3845.

APPENDICES

A1 COVID-19 Cases

Table 1. Daily counts of infected- and deceased cases in New York (CDC, 2022)

Date	Total infected	Total Deaths
3/1/2020	0	0
3/2/2020	0	0
3/3/2020	0	0
3/4/2020	1	0
3/5/2020	12	0
3/6/2020	21	0
3/7/2020	28	0
3/8/2020	28	0
3/9/2020	125	0
3/10/2020	137	0
3/11/2020	164	0
3/12/2020	230	0
3/13/2020	267	0
3/14/2020	311	0
3/15/2020	400	1
3/16/2020	487	1
3/17/2020	730	1
3/18/2020	1043	1
3/19/2020	1683	5

continued on next page

Table 1. *continued*

Date	Total infected	Total Deaths
3/20/2020	2694	15
3/21/2020	4145	21
3/22/2020	6123	24
3/23/2020	8570	31
3/24/2020	10761	39
3/25/2020	12955	52
3/26/2020	15865	59
3/27/2020	19237	96
3/28/2020	22552	107
3/29/2020	25745	119
3/30/2020	29044	155
3/31/2020	32656	211
4/1/2020	36273	275
4/2/2020	40572	395
4/3/2020	45704	508
4/4/2020	50398	566
4/5/2020	54480	1017
4/6/2020	58188	1197
4/7/2020	61897	1378
4/8/2020	67513	1545
4/9/2020	72910	1787
4/10/2020	78128	2024
4/11/2020	82150	2260

continued on next page

Table 1. *continued*

Date	Total infected	Total Deaths
4/12/2020	85486	2443
4/13/2020	88268	2661
4/14/2020	91743	2882
4/15/2020	95477	3081
4/16/2020	99138	3247
4/17/2020	102290	3560
4/18/2020	105548	3562
4/19/2020	108350	3787
4/20/2020	110706	3940
4/21/2020	112365	4107
4/22/2020	114784	4260
4/23/2020	117605	4407
4/24/2020	121117	4552
4/25/2020	127030	4715
4/26/2020	129787	4831
4/27/2020	131507	4947
4/28/2020	132768	5060
4/29/2020	134850	5170
4/30/2020	136894	5345
5/1/2020	138624	5442
5/2/2020	140623	5544
5/3/2020	142084	5651
5/4/2020	143302	5808

continued on next page

Table 1. *continued*

Date	Total infected	Total Deaths
5/5/2020	144318	5907
5/6/2020	145627	6482
5/7/2020	147253	6580
5/8/2020	148624	6657
5/9/2020	149833	6766
5/10/2020	150978	6867
5/11/2020	151698	6947
5/12/2020	152362	7045
5/13/2020	153411	7132
5/14/2020	154506	7211
5/15/2020	155456	7279
5/16/2020	156632	7377
5/17/2020	157528	7448
5/18/2020	158141	7496
5/19/2020	159024	7550
5/20/2020	159820	7606
5/21/2020	160783	7660
5/22/2020	161670	7716
5/23/2020	162660	7762
5/24/2020	163392	7830
5/25/2020	164033	7879
5/26/2020	164535	7927
5/27/2020	164997	7977

continued on next page

Table 1. *continued*

Date	Total infected	Total Deaths
5/28/2020	165682	8023
5/29/2020	166285	8058
5/30/2020	166909	8100
5/31/2020	167467	8130
6/1/2020	167947	8159
6/2/2020	168663	8198
6/3/2020	169213	8230
6/4/2020	169727	8259
6/5/2020	170268	8284
6/6/2020	170805	8308
6/7/2020	171128	8339
6/8/2020	171446	8362
6/9/2020	171789	8391
6/10/2020	172038	8416
6/11/2020	172375	8438
6/12/2020	172760	8468
6/13/2020	173137	8489
6/14/2020	173446	8502
6/15/2020	173685	8521
6/16/2020	173984	8538
6/17/2020	174201	8551
6/18/2020	174500	8568
6/19/2020	174886	8580

continued on next page

Table 1. *continued*

Date	Total infected	Total Deaths
6/20/2020	175211	8595
6/21/2020	175490	8606
6/22/2020	175747	8613
6/23/2020	176029	8627
6/24/2020	176318	8636
6/25/2020	176716	8645
6/26/2020	177150	8654
6/27/2020	177489	8664
6/28/2020	177789	8667
6/29/2020	177991	8673
6/30/2020	178275	8680
	...	
5/1/2021	1121015	19409
5/2/2021	1122872	19431
5/3/2021	1124156	19454
5/4/2021	1125497	19480
5/5/2021	1126855	19499
5/6/2021	1128398	19508
5/7/2021	1129878	19522
5/8/2021	1131817	19540
5/9/2021	1133242	19551
5/10/2021	1134154	19561
5/11/2021	1135113	19581

continued on next page

Table 1. *continued*

Date	Total infected	Total Deaths
5/12/2021	1136256	19591
5/13/2021	1137617	19600
5/14/2021	1138883	19612
5/15/2021	1140253	19623
5/16/2021	1141186	19638
5/17/2021	1141984	19645
5/18/2021	1142612	19654
5/19/2021	1143571	19666
5/20/2021	1144584	19680
5/21/2021	1145551	19688
5/22/2021	1146403	19702
5/23/2021	1147037	19706
5/24/2021	1147626	19715
5/25/2021	1148105	19722
5/26/2021	1148695	19736
5/27/2021	1149326	19742
5/28/2021	1149873	19753
5/29/2021	1150405	19759
5/30/2021	1150933	19770
5/31/2021	1151220	19780
6/1/2021	1151435	19788
6/2/2021	1151698	19792
6/3/2021	1151994	19799

continued on next page

Table 1. *continued*

Date	Total infected	Total Deaths
6/4/2021	1152423	19808
6/5/2021	1152913	19818
6/6/2021	1153222	19830
6/7/2021	1153515	19835
6/8/2021	1153736	19844
6/9/2021	1153952	19848
6/10/2021	1154216	19852
6/11/2021	1154481	19858
6/12/2021	1154753	19867
6/13/2021	1154953	19872
6/14/2021	1155117	19878
6/15/2021	1155263	19884
6/16/2021	1155433	19887
6/17/2021	1155645	19889
6/18/2021	1155881	19889
6/19/2021	1156118	19902
6/20/2021	1156172	19902
6/21/2021	1156227	19906
6/22/2021	1156354	19910
6/23/2021	1156486	19916
6/24/2021	1156655	19916
6/25/2021	1156796	19919
6/26/2021	1156956	19922

continued on next page

Table 1. *continued*

Date	Total infected	Total Deaths
6/27/2021	1157104	19925
6/28/2021	1157238	19927
6/29/2021	1157344	19928
6/30/2021	1157475	19930
7/1/2021	1157650	19932
7/2/2021	1157893	19933
7/3/2021	1158044	19933
7/4/2021	1158205	19935
7/5/2021	1158346	19936
7/6/2021	1158474	19939
7/7/2021	1158647	19942
7/8/2021	1158865	19944
7/9/2021	1159186	19947
7/10/2021	1159441	19947
7/11/2021	1159695	19947
7/12/2021	1159918	19947
7/13/2021	1160203	19952
7/14/2021	1160539	19954
7/15/2021	1160945	19955
7/16/2021	1161322	19956
7/17/2021	1161724	19957
7/18/2021	1162215	19958
7/19/2021	1162577	19959

continued on next page

Table 1. *continued*

Date	Total infected	Total Deaths
7/20/2021	1163052	19960
7/21/2021	1163634	19963
7/22/2021	1164284	19964
7/23/2021	1165058	19966
7/24/2021	1165798	19966
7/25/2021	1166710	19973
7/26/2021	1167263	19975
7/27/2021	1168038	19976
7/28/2021	1168983	19981
7/29/2021	1170157	19981
7/30/2021	1171416	19983
7/31/2021	1172784	19985
8/1/2021	1174038	19990
8/2/2021	1174979	19990
8/3/2021	1176467	19991
8/4/2021	1177993	19996
8/5/2021	1179682	20000
8/6/2021	1181447	20002
8/7/2021	1183888	20009
8/8/2021	1185661	20014
8/9/2021	1187342	20020
8/10/2021	1189064	20028
8/11/2021	1191255	20039

continued on next page

Table 1. *continued*

Date	Total infected	Total Deaths
8/12/2021	1193699	20052
8/13/2021	1196273	20060
8/14/2021	1198595	20069
8/15/2021	1200816	20080
8/16/2021	1202885	20086
8/17/2021	1204930	20096
8/18/2021	1207588	20111
8/19/2021	1210346	20123
8/20/2021	1213128	20135
8/21/2021	1216034	20148
8/22/2021	1218532	20164
8/23/2021	1220826	20177
8/24/2021	1222592	20185
8/25/2021	1224985	20196
8/26/2021	1228157	20212
8/27/2021	1232380	20226
8/28/2021	1235297	20248
8/29/2021	1238214	20270
8/30/2021	1240856	20278
8/31/2021	1243236	20290
9/1/2021	1246065	20308
9/2/2021	1249321	20332
9/3/2021	1253112	20352

continued on next page

Table 1. *continued*

Date	Total infected	Total Deaths
9/4/2021	1256486	20370
9/5/2021	1259355	20392
9/6/2021	1261814	20410
9/7/2021	1264076	20436
9/8/2021	1266531	20456
9/9/2021	1269961	20471
9/10/2021	1274025	20498
9/11/2021	1277616	20518
9/12/2021	1281351	20538
9/13/2021	1283649	20555
9/14/2021	1286202	20576
9/15/2021	1289383	20601
9/16/2021	1294121	20628
9/17/2021	1297994	20651
9/18/2021	1301280	20670
9/19/2021	1304800	20695
9/20/2021	1306963	20716
9/21/2021	1310574	20745
9/22/2021	1313543	20770
9/23/2021	1316837	20799
9/24/2021	1321300	20818
9/25/2021	1325098	20842
9/26/2021	1328061	20871

continued on next page

Table 1. *continued*

Date	Total infected	Total Deaths
9/27/2021	1330887	20887
9/28/2021	1333608	20908
9/29/2021	1337434	20938
9/30/2021	1340812	20968
10/1/2021	1344740	20991
10/2/2021	1348076	21009
10/3/2021	1351472	21033
10/4/2021	1353306	21047
10/5/2021	1356030	21075
10/6/2021	1360301	21098
10/7/2021	1363979	21118
10/8/2021	1368316	21151
10/9/2021	1371855	21179
10/10/2021	1373932	21204
10/11/2021	1377715	21226
10/12/2021	1380063	21251
10/13/2021	1383582	21276
10/14/2021	1386927	21314
10/15/2021	1391234	21336
10/16/2021	1395252	21355
10/17/2021	1398401	21386
10/18/2021	1400717	21407
10/19/2021	1402567	21432

continued on next page

Table 1. *continued*

Date	Total infected	Total Deaths
10/20/2021	1406104	21458
10/21/2021	1409282	21498
10/22/2021	1412017	21518
10/23/2021	1415481	21549
10/24/2021	1417633	21573
10/25/2021	1419988	21594
10/26/2021	1421629	21626
10/27/2021	1425129	21656
10/28/2021	1428328	21680
10/29/2021	1432516	21707
10/30/2021	1435659	21734
10/31/2021	1438453	21760
11/1/2021	1440632	21786
11/2/2021	1442927	21800
11/3/2021	1446157	21513
11/4/2021	1449814	21542
11/5/2021	1453696	21570
11/6/2021	1457547	21604
11/7/2021	1461262	21624
11/8/2021	1463956	21646
11/9/2021	1467214	21676
11/10/2021	1471378	21701
11/11/2021	1476679	21735

continued on next page

Table 1. *continued*

Date	Total infected	Total Deaths
11/12/2021	1481688	21760
11/13/2021	1486783	21775
11/14/2021	1490979	21796
11/15/2021	1494504	21820
11/16/2021	1498463	21846
11/17/2021	1503387	21867
11/18/2021	1509904	21894
11/19/2021	1515921	21925
11/20/2021	1521477	21943
11/21/2021	1526969	21967
11/22/2021	1531063	21994
11/23/2021	1535193	22018
11/24/2021	1540890	22043
11/25/2021	1547652	22067
11/26/2021	1552589	22067
11/27/2021	1555836	22123
11/28/2021	1560342	22150
11/29/2021	1564336	22185
11/30/2021	1569506	22221

A2 Pseudo Code: PRecomm

Algorithm 1 PRecomm (Prediction & Recommendation)

procedure INITIALIZATION
    ny ←tweets in New York (after pre-processing)
    ny[y] ←[1, 2, 3, 4, 5] (encoded values for each preventive behavior)
    un_ny ←NOT encoded tweets only
    cnt ←0
    n ←cosine distance (knn parameter)
    y ←prediction from svm ([prevention behaviors, probabilities])
    y2 ←prediction from knn ([prevention behaviors, similarity scores])
procedure INITIAL- & SELF-TRAINING
    if (cnt < 3) **then**:
        if (ny[y] **is not null**) **then**
            TfidfVectorizer(ny[body], ny[location], ny[bio])
            [train, test] ←split_train_test(ny, 8:2)
            for (c in [0.1, 0.2, 0.7, 0.9, 1.0, 1.8, 2.0]) **do**
                svm(c, penalty) with train, train[y] (Train the svm model)
                test[yhat] ←Predict test[y] by the trained svm
                f1_score ← Append f1_score(test[y], test[yhat])
            c ←Select c with the highest f1_score
            for (n in [1, 2, 3, ..., 20]) **do**
                knn(n, cosine) with train, train[y] (Train the knn model)
                test[yhat] ←Predict test[y] by the trained **knn**(n, cosine)
                f1_score ←Append f1_score(test[y], test[yhat])
            n ←Select n with the highest f1_score
        else:
            `remove the null data`
        procedure PREDICTION & RECOMMENDATION
            [ny[y], probability] ←**svm**(un_ny)
            GO TO 12:
            [ny[y2], similarity] ←**knn**(un_ny)
            GO TO 20:
            cnt = cnt+1
        return ny[y, y2], probability, similarity
    else:
        stop