Designing An Efficient Public Health Surveillance System Using Machine Learning

Thesis submitted in partial fulfilment of the requirements for the award of the degree of

Doctor of Philosophy

by

AAKANSHA GUPTA

(2K18/PhD/CO/10)

Under the supervision of

Prof. Rahul Katarya

Department of Computer Science and Engineering

Delhi Technological University

Delhi, India

2022

Dedicated to

My beloved Parents

CANDIDATE DECLARATION

I hereby declare that the thesis entitled "Designing An Efficient Public Health Surveillance System Using Machine Learning" submitted to Delhi Technological University, Delhi, in the partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy in the Department of Computer Science, is an original work and has been done by myself under the supervision of Prof. Rahul Katarya (Supervisor), Department of Computer Science and Engineering, Delhi Technological University, Delhi, India.

The interpretations presented are based on my study and understanding of the original texts. The work reported here has not been submitted to any other institute for the award of any other degree.

Aakansha Gupta
Roll No. 2K18/PhD/CO/10
Department of Computer Science and Engineering
Delhi Technological University
Delhi-110042, India

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
(Govt. of National Capital Territory of Delhi)
Shahbad Daulatpur, Main Bawana Road,
Delhi-110042, India

Date:_____

CERTIFICATE

This is to certify that the work incorporated in the thesis entitled "Designing An Efficient Public Health Surveillance System Using Machine Learning" submitted by Ms. Aakansha Gupta (Roll No. 2K18/PhD/CO/10) in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy, to the Delhi Technological University, Delhi, India is carried out by the candidate under my supervision and guidance at the Department of Computer Science and Engineering, Delhi Technological University, Delhi, India.

The results embodied in this thesis have not been presented to any other University or Institute for the award of any degree or diploma.

Prof. Rahul Katarya
Department of Computer Science and Engineering
Delhi Technological University
Delhi-110042, India

ACKNOWLEDGMENT

I address my sincere thanks to Almighty God for giving me the inner power to complete my thesis and guide me in every step of my life.

It is an immense pleasure to have the opportunity to express my heartfelt gratitude to everyone who helped me throughout this research. I would like to express my heartfelt gratitude and indebtedness to my supervisor Prof. Rahul Katarya (Dept. of Computer Science & Engineering), for his invaluable and positive guidance, encouragement, and patience. During the research, his motivation and encouragement have inspired me to grow as a scholar and as a person. I am deeply indebted to my supervisor for guiding me in carrying out the research work and morally supporting me in every way during the course's challenging times. His technical expertise, precise suggestions, kind nature, and detailed, timely discussions are wholeheartedly appreciated. Also, my sincere thank goes to Delhi Technological University for considering my candidature for this course. I am also very thankful to Prof. Jai Prakash Saini, Vice-Chancellor, Delhi Technological University, Delhi, India, who has been a constant source of enthusiasm. He has always motivated young researchers like me to pursue excellence to achieve higher goals in academics and research. Also, my sincere thanks reciprocate to Dr. Vinod Kumar (HoD, Dept. of Computer Science and Engineering), Prof. Rajni Jindal (Chairperson DRC, Dept. of Computer Science and Engineering) for insightful comments and valuable suggestions. Special thanks to my seniors and colleagues of Delhi Technological University, Delhi, India. My sincere thanks to all the professors, faculty, researchers, and nonteaching staff of the Computer Science Department. I would also like to express my gratitude to the University Grants Commission (UGC), New Delhi, for providing financial support for the study in the form of Junior Research Fellowship (JRF) and Senior Research Fellowship (SRF).

I also wish to take this opportunity to thank all my teachers who have taught me and shaped me into the person I am, aggravated me to be an academician, and have directly indirectly made me capable of succeeding in completing this research work. I am thankful to all my colleagues and friends during my journey as a Ph.D. scholar. The engaging talks, brainstorming, and collaborative teamwork significantly impacted my growth as an independent researcher.

Finally, but most importantly, my heartfelt gratitude is for my parents, who are the motivations behind me; without their blessings, this work could not have been accomplished. I want to thank my husband, sister, brother, and family, especially my Nani Ji, for their endless love, support, and encouragement. I am truly indebted to them.

Aakansha Gupta
(2K18/PhD/CO/10)
Department of Computer Science
Delhi Technological University,
Delhi-110042, India

ABSTRACT

Public Health Surveillance (PHS) is considered to be an essential public health function. The primary functions of a public health system are health surveillance, population health assessment, disease and injury prevention, health protection, and health promotion. Surveillance is defined as “the close and continuous monitoring of one or more people for the purpose of direction, supervision, or control”. The World Health Organization (WHO) defines public health surveillance as “the ongoing, systematic collection, analysis, and interpretation of health-related data essential to the planning, implementation, and evaluation of public health practice”. Public health surveillance is regarded as the most effective tool for preventing epidemics. Health Surveillance can be used to track chronic diseases, infectious diseases, injuries, healthcare utilization, environmental concerns, and vector dispersal. Surveillance data are essential for influencing policy decisions, leading new program activities, refining public communications, and aiding agencies in evaluating research investments. Addressing case under-ascertainment is important in most surveillance systems, especially in pandemics of novel diseases with a wide range of clinical presentations, because it might impact public risk perception and policy implementation time. However, surveillance is never perfect, and diseases with a high proportion of mild, pauci-symptomatic, or subclinical cases can be difficult to identify and contain in most indicator-based monitoring systems.

Effective public health surveillance systems can provide timely and reliable information allowing for the early detection of potential epidemics. A systematic approach is required to strengthen public health surveillance systems that can quickly detect and respond to the initial cases of disease outbreaks and other public health emergencies. One of the primary purposes of public health surveillance is to monitor diseases and trends of public health events to ensure that any atypical disease patterns, such as outbreaks, are timely discovered, examined, and responded to. Incentives are in place to encourage the development of public health surveillance systems, and employing machine learning technologies in public health events can help public health professionals speed up the process of monitoring, evaluating, and decision-making. With the increase in the use of the internet, the digital world is generating data at an alarming and continuing rate. One approach is to leverage the online health mentions posted during an ongoing

public health event that generates unprecedented amounts of health-related data and couple it with the modern Machine Learning techniques for decision support. As appealing as it may sound, incorporating online data is associated with data-science challenges that limit the effective learning of ML models. The primary focus of this thesis is to incorporate public health-related social media data along with historical data to improve prediction performance. The proposed models are empirically evaluated in the context of predicting health events such as novel COVID-19 disease cases. Overall, the research work is primarily useful for tracking and forecasting an ongoing outbreak, and it can give valuable advice to disease makers and epidemiologists. As a result, they will be able to implement appropriate policies to prevent and manage the epidemic. Therefore, this study represents an organized, systematic, and arranged effort that determines the identification, power, and applicability of public health surveillance using machine learning techniques.

Objectives: The objectives of the entire study have been classified into four segments:

- The first objective of the study is to enhance the prediction performance of the epidemiological models based on Machine Learning (ML) techniques.
- The second objective focuses on analyzing the impact of the health determinant factors such as demographic data, environmental data, etc., and their significance for public health surveillance.
- The third objective is to improve the feature extraction to enhance the performance of ML algorithms.
- The last objective is to explore the applicability of simultaneously using multiple online platforms to improve prediction accuracy.

Methodology: For achieving the mentioned objectives, this study utilizes machine learning and deep learning techniques like evolutionary algorithms, Neural Networks (NN), Natural Language Processing (NLP), and Topic Modeling approaches due to the tremendous applicability to solving the natural world problems. The following strategies are used to achieve the targeted objectives:

- For achieving the first objective, innovative and novel models based on mathematical statistics, machine learning, and deep learning have been used to predict epidemic time

series. The NLP techniques, Pandemic-Latent Dirichlet allocation (PAN-LDA) based Long Short-Term Memory (LSTM) neural network, and evolutionary algorithm were used to improve the prediction accuracy.

- To accomplish the second objective, we analyzed the pandemic-related multi-source data and their impact on disseminating the pandemic. Later, we presented a prediction model that incorporates the health determinants data with the historical cases to predict the outbreak.
- To attain the third objective, we proposed two feature extraction algorithms. One algorithm employed semantically and morphologically similar word embedding clusters as features to improve the clustering performance. And the other algorithm used Natural Language Processing and Topic Modeling approaches by incorporating historical cases and the corresponding news articles to extract better features for the time-series prediction.
- The last objective explored the relevance and applicability of multi-source internet data to enhance prediction performance. With the increase in the use of online platforms, there is a tremendous increase in the data posted about the ongoing events. This large amount of data from multiple social media platforms can improve the performance of epidemic models.

Results: The outcomes of the study are as follows:

- An evolutionary algorithm and LSTM-based epidemic model is proposed to perform epidemic trend predictions.
- A study is conducted to geospatially analyze the demographic, health, socio-economic, and climatic factors associated with the pandemic distribution.
- A fixed-effect multiple regression prediction model is proposed to predict the daily confirmed cases during the early phases of the COVID-19 second wave and determine the possibility of upcoming waves.
- A document representation method is proposed based on semantically and morphologically similar feature clusters to enhance the clustering performance.

- A Latent Dirichlet Allocation (LDA) based PAN-LDA feature extraction model is developed that makes use of the historical cases and the corresponding news articles to extract better features for the prediction.
- A study was performed to analyze the trends in social media-based public health surveillance systems using ML algorithms.
- An epidemic model is proposed that incorporates the features from data collected from multiple online sources such as Twitter, Reddit, and Google news to improve the prediction performance.

TABLE OF CONTENTS

CHAPTER 5: VECTOR REPRESENTATION OF DOCUMENTS USING FEATURE CLUSTERS

LIST OF ABBREVIATIONS

ACC	Accuracy
AE	Autoencoder
AE	Autoencoder
AI	Artificial Intelligence
ANN	Artificial Neural Network
API	Application Programming Interface
ARI	Adjusted Rand Index
BERT	Bidirectional Encoder Representations from Transformers
Bi-LSTM	Bidirectional Long Short-Term Memory
CART	Classification And Regression Trees
CBOW	Continuous Bag-of-Words
CDC	Centers for Disease Control and Prevention
CGS	collapsed Gibbs sampling
CMR	Community Mobility Reports
CNN	Convolution Neural Network
COVID-19	Coronavirus Disease-2019
DEC	Deep Embedded Clustering
DEC	Deep Embedded Clustering
DL	Deep Learning
DNN	Deep Neural Network
DT	Decision Tree
ES	Evolutionary Strategies
FE_MR	Fixed-Effect Multiple Regression
GMI	Global Moran's Index
GN	Google News
GPM	Grocery and Pharmacy Mobility
GT	Google Trends

HB	Hospital Beds
HH	High-High
HL	High-Low
idf	inverse document frequency
IDSP	Integrated Disease Surveillance Program
ILI	Influenza-Like Illness
KL	Kullback–Leibler
k-NN	k-Nearest Neighbor
KPCA	Kernel Principal Component Analysis
LASSO	Least Absolute Shrinkage And Selection Operator
LDA	Latent Dirichlet allocation
LE	Life Expectancy
LH	Low-High
LightGBM	Light Gradient Boosting Machines
LL	Low-Low
LMI	Local Moran's Index
LR	Logistic Regression
LSTM	Long Short-term Memory
MAD	Mean Absolute Deviation
MSE	Mean Square Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MERS	Middle East Respiratory Syndrome
MI	Mutual information
ML	Machine Learning
MLR	Multiple Linear Regression
NB	Naive Bayes
NBM	Multinomial Naive Bayes
NEDSS	National Electronic Disease Surveillance System
NETSS	National Electronic Telecommunications Systems for Surveillance

NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NMI	Normalised Mutual Information
NN	Neural Network
NPI	Non-Pharmaceutical Interventions
PAN-LDA	Pandemic-LDA
PD	Population Density
PHS	Public Health Surveillance
PLSI	Probabilistic Latent Semantic Indexing
PR	Poverty Rate
R^2	R-Square
RD	Reddit
RF	Random Forest
RI	Rand Index
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
RoBERTa	Robustly optimized BERT approach
RR	Ridge Regression
RRM	Retail and Recreation Mobility
SAE	Stacked Autoencoder
SEIR	Susceptible-Exposed-Infected-Recovered
SGNS	Skip-Gram with Negative Sampling
SIQRV	Susceptible–Infected–Quarantined–Recovered–Vaccinated
SIR	Susceptible-Infected-Recovered
SIRD	Susceptible -Infected-Recovered-Dead
SVM	Support Vector Machine
tf	term frequency
TW	Twitter
UP	Urban Population
VIF	Variance Inflation Factor

VSM	Vector Space Models
WC_MTI	Word Cluster-based Modified Tf-Idf
WHO	World Health Organization
XGBoost	Extreme Gradient Boosting

LIST OF TABLES

LIST OF FIGURES

Chapter 1 INTRODUCTION

The emerging field of Public Health Surveillance (PHS) offers promising tools that can assist public health officials in streamlining the process of monitoring, analyzing, and utilizing unofficial sources to help in decision-making. Public Health Surveillance is the foundation of public health practice. Surveillance data are essential for influencing policy decisions, leading new program activities, refining public communications, and aiding agencies in evaluating research investments [1]. Surveillance is depicted as a vital backbone for the discipline as a whole by highlighting its ability to feed public health initiatives and decision-making processes. Public health surveillance is distinct from most modern surveillance approaches in that it focuses on health problems, diseases, and contagions rather than individuals. Surveillance has evolved significantly over the years and will most certainly continue to do so. Changes in disease monitoring systems have recently occurred as a result of technological breakthroughs in data gathering connected to Internet availability and increased processing capability. Digital data has mostly been exploited to predict or nowcast infectious disease epidemics. Surveillance operations frequently employ Machine Learning (ML) techniques because of the abundance and variety of data available via digital sources. The development of statistical and machine learning methodologies with a focus on clinical tasks, such as forecasting disease prognosis and detecting phenotypes, has substantially evolved in the last decade to the benefit of patients. In addition, the application of Machine Learning techniques in healthcare is increasing, which enables more effective estimation and prediction of health outcomes from massive data sets. Section 1.1 covers a detailed view of public health surveillance along with its objectives. Also, the significance of internet technology in public health surveillance is discussed in detail. Section 1.2 illustrates the overview of machine learning techniques and their applications in the healthcare sector. Section 1.3 describes public health surveillance using machine learning techniques. Section 1.4 covers the motivation behind the study. Section 1.5 covers the research questions and research objectives in detail. The brief of all the chapters is discussed in Section 1.6. Finally, section 1.7 wraps the chapter with a summary of the entire chapter.

1.1. Public Health Surveillance (PHS)

Real-time surveillance in the field of health informatics has emerged as a growing domain of interest among worldwide researchers. Evolution in this field has helped in the introduction of various initiatives related to public health informatics. The term "surveillance" is derived from the French words sur (over) and veiller (to watch). Globally, there is a recognition of the need to enhance disease surveillance and response systems. A well-functioning disease surveillance system offers data for public health

intervention program development, implementation, monitoring, and assessment. According to the World Health Organisation (WHO), Public Health Surveillance is "An ongoing, systematic collection, analysis and interpretation of health-related data essential to the planning, implementation, and evaluation of public health practice" [2] (Figure 1.1). Surveillance is crucial for assessing the necessity for interventions as well as the effectiveness of treatments since it can directly examine what is happening in the community. Infectious disease monitoring is acknowledged as a critical component of public health decision-making and practice. Surveillance aims to give decision-makers timely, meaningful evidence, allowing them to lead and manage more effectively.

Information and data from PHS are utilized to evaluate and categorize the strain and allocation of severe health events, prioritize public health interventions, monitor the efficacy of the control measures, and detect new medical problems that may substantially impact society. Surveillance statistics are critical for monitoring the population's health state, diagnosing diseases, and initiating action to avoid additional diseases and contain public health concerns. The primary purpose of surveillance systems in public health practice and their ability to significantly impact the efficacy and effectiveness of the public health system have motivated research to strengthen the scientific base of PHS [3]. Early warning of epidemics is critical for effective and speedy control, and information on endemic communicable diseases is critical for disease monitoring. Many nations have built surveillance technology to monitor high-burden diseases and detect outbreaks of epidemic-prone diseases.

PHS extends back to the reign of Pharaoh Mempses in the First Dynasty when the first pandemic in human history was documented [4]. Many foreboding signs and a huge plague happened during his rule" [5]. The "great plague" is now thought to have taken place around 3180 B.C. [6]. Hippocrates (460 B.C.–370 B.C.) [7], a Greek physician recognized as the "father of medicine" is credited with the concept of data collection and analysis.

The notion of PHS has developed through time and is still mixed up with other definitions of the term. Dr. Alexander D. Langmuir proposed the present notion of surveillance as the monitoring of disease incidence in populations as a function of the newly formed Communicable Disease Center (now the Centers for Disease Control and Prevention (CDC)) [8]. Previously, surveillance was used to closely monitor persons infected with a communicable disease to discover early signs and symptoms and undertake timely quarantine and control measures. In order to differentiate between these two forms of surveillance, the term PHS is now used to refer to the tracking of health events in the population, and medical surveillance refers to the surveillance of possibly exposed patients to identify initial symptoms.

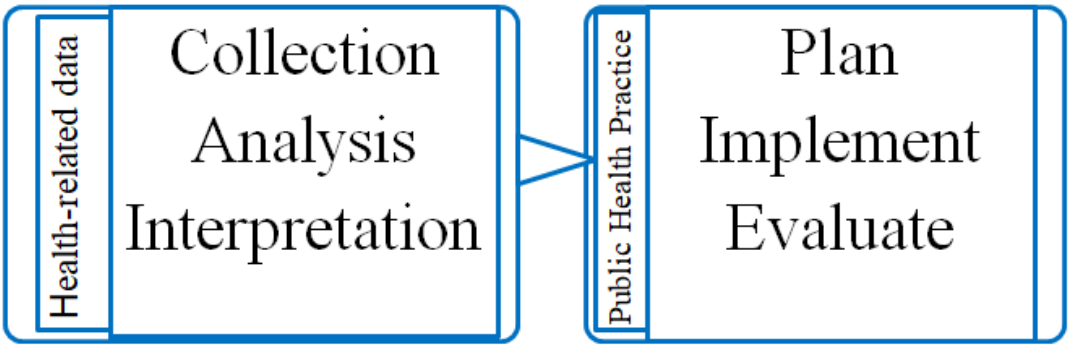


Figure 1.1 Steps to Surveillance System

1.1.1. Objectives of Surveillance Systems

PHS offers the empirical and factual information required to make informed decisions and implement successful public health interventions. The primary goal of surveillance is to provide data that can guide actions. The design and execution of surveillance systems are driven by the public health objectives and actions required for successful intervention. For example, if the goal is to prevent the spread of acute infectious disease outbreaks, management must act quickly to minimize disease transmission. As a result, a monitoring system that provides timely early warning information from laboratories and clinics is required. Chronic diseases and health-related habits, on the other hand, evolve slowly. A surveillance system to monitor the population effects of a tuberculosis control program, for example, may provide information only once every one to five years via a series of demographic and health surveys [9]. The concept is that various public health objectives and actions required to attain them need various information systems. The type of surveillance or health information system that should be used should be defined by the type of action that can be taken, as well as when and how frequently it must be performed, and what information or data is required for its execution.

In order to characterize the activities of a surveillance system, the following questions must be answered:

- What is the case definition of the health event? Is it practical in this situation?
- What is the purpose of surveillance, and what are its goals?
- Does the system integrate with other surveillance and health information systems?
- How is data handled? How are they transmitted, routed, and stored? Are there any needless delays? How is confidentiality ensured?
- What are the data sources? Who is supposed to report? Who actually does report?
- What information is gathered? Is it what programs require?
- What is the frequency of data collection (weekly, monthly, yearly)?

- How are the data analyzed? By whom? How frequently? How thoroughly?
- How is the information spread? How frequently are reports distributed? To whom? Is it reaching all those who need to know, such as the medical and public health communities as well as policymakers?
- What are the planned uses of the surveillance data?
- What is the population under surveillance?

Finally, public health surveillance aims to study the continuous trend of disease incidence and disease potential in a community for effective research, management, and prevention of disease. Generally, public health officials typically react to reports of contagious diseases by implementing basic control methods such as quarantine. Agencies may now utilize surveillance data to develop more efficient disease control and preventive actions. However, public health surveillance is not limited to diseases that have efficient control methods. The surveillance can also be justified for two more reasons: for beginners, surveillance allows us to understand more about a disease's evolutionary biology, medical spectrum, and demography. This understanding might lead to the development of preventative and control methods. Second, surveillance gives a foundation of data to develop and implement preventative and control strategies.

1.1.2. Principles and Uses of Surveillance

According to Foege, Hogan, and Newton, the goal of collecting, evaluating, and distributing disease information is to control that disease [10]. The fundamental principle of PHS is that it should be designed and implemented to give accurate information to policymakers in a timely and cost-effective manner. Since management is unlikely to take actions to address slight variations between regions, forgoing accuracy makes sense to enhance the timeliness and conserve resources that might be utilized for public health actions. The utilization of surveillance data may be considered instant, yearly, or historical, based on the public health interventions that may be carried out.

The World Bank classified public health surveillance into six groups [11] :

- Recognize cases or clusters of cases to initiate actions to prevent transmission or lower morbidity and death.
- Determine and measure patterns, or assess the public health effect of health occurrences.
- Examine the efficacy of preventive and control measures, as well as intervention options.
- Demonstrate the importance of public health intervention programs and resources, and allocate resources during public health planning.
- Determine high-risk demographic groups or geographic regions for intervention and drive analytic research.

- Create hypotheses that will lead to analytic investigations on disease risk factors for cause, propagation, or progression.

1.1.3. Sources of Data

There are several data sources accessible for public health surveillance. Some of the most commonly used sources of surveillance data are [12] :

- Reports of epidemic investigations
- Reports of individual case investigations
- Reports of laboratory utilization
- Epidemic reports
- Mortality reports
- Morbidity reports
- Special surveys
- Environmental data
- Demographic data
- Information on animal reservoirs and vectors
- Sentinel surveillance
- Syndromic data
- National statistics
- Social media

1.1.4.Enhancing the Use of Computer Technology in Public Health Surveillance

Public health monitoring is based on public health information systems designed to encompass a diverse set of data sources vital to public health action [13]. Computer technologies have the potential to enhance public health information systems, which vary from basic systems that collect data from a particular source to digital systems that take data in a variety of forms and complex surveys. Despite its shortcomings [14], the use of digital technology continues to advance public health monitoring [15]. For example, in the United States of America, by 1991, the National Electronic Telecommunications Systems for Surveillance (NETSS) had computerized all of the nation's health departments for the systematic gathering, processing, and sharing of information on reportable conditions [16]. Later, CDC implemented the National Electronic Disease Surveillance System (NEDSS) in order to effectively enhance and control the existing surveillance systems and give people the opportunity to respond to health hazards quite quickly [17]. Public health specialists and government organizations would be able to identify and

respond to epidemics and bioterrorism strikes in real-time once the monitoring system is fully implemented across the nation.

Computers and data have altered the way we organize information, communicate, and think about science. The curation of massive datasets has transformed numerous fields, allowing breakthroughs on a scale previously unimaginable. However, data in isolation has little relevance; our efforts should aim to obtain actionable insights and knowledge that may be used to influence decisions and eventually enhance people's lives.

There is a vigorous debate about how new computer technology has the potential to increase the quality, capability, and efficacy of public health monitoring systems. An example is the usage of "e-Health," which is a prospective dynamic health information system. e-Health is a comparatively new word for healthcare practice aided by electronic procedures and communication [18]. Another technique is a novel method for detecting outbreaks using search engine data [19]. To examine the highly infectious disease, measles, in different countries of Europe, researchers evaluated data obtained from Google Trends (GT) for the duration of 5 years and achieved an early prediction at least 2 months in advance. Another example is a recent study of how online surveillance techniques might aid in the early detection of epidemics [20]. The study discovered that online information sources enable faster identification of epidemics, lower costs, and enhance reporting openness. As the quantity and diversity of systems grow, future public health monitoring initiatives should emphasize developments in electronic data exchange and data integration, as well as patient confidentiality, data protection, and system security [21].

1.1.5. Public Health Surveillance and Internet Technology

The ubiquitous and publicly available information created on the Internet has spurred a growing interest in establishing digital public health monitoring systems. New terms such as "infodemiology" and "infoveillance" have been introduced to describe the use of informatics methodologies to analyze information from the Internet in order to forecast epidemics [22]. Infodemiology is the science of distribution and sources of information in an electronic medium to inform public policy and public health. Examples of infodemiology applications include: analyzing Internet search engine queries to forecast epidemics, classifying and analyzing public health-related online articles, monitoring users' posts on online platforms such as Twitter (TW) for syndromic surveillance, and identifying and assessing inconsistencies in information availability. Using infodemiology data for surveillance purposes has been called "infoveillance".

The Internet is now ingrained in almost everyone's everyday lives. The availability of online data sources has developed in recent years as an addition to conventional surveillance methods [23] and has considerably contributed to contagious disease monitoring by delivering real-time information and cutting

public health expenses [24]. When a user's social media post results in a large increase in people's participation in skin cancer prevention, it is worth emphasizing how rapidly incidences may be identified in real-time using infodemiology data [25]. Regardless of the varied applications of social media data, the impact of monitoring social media in making healthcare decisions might be studied [26]. Furthermore, the demand for novel methodology and techniques to improve the capabilities of existing syndromic monitoring systems has arisen as a result of early disease detection and prompt public health intervention. Several research papers [24,27–31] have been published to evaluate and design surveillance systems that use social media data. Applications, technology, methodologies, data sources, and evaluation were all covered in these researches.

Recent advancements in social media platforms, as well as their widespread usage, have opened new opportunities for gathering health and well-being data in a continuous manner outside of laboratory and hospital settings. In addition to "filling gaps" in conventional clinical data, these platforms bring up new scientific and business opportunities for large-scale lifestyle monitoring. Simultaneously, apparently unrelated variables such as mature open-source scientific software libraries, faster data crowdsourcing and labeling, and the reuse of specialized hardware have permitted remarkable advances in predictive modeling. Many ML projects have shown outstanding results, ranging from object detection to professionals in breast cancer screening. In all of these cases, the common factor has been the management of huge high-quality datasets that allow models to explore hidden patterns and then generalize in the real world. However, particularly in health care, where inaccurate predictions can have serious implications, the roll-out and acceptance of such systems have been faced with criticism. Instead, fields with low false-positive costs and increased digitization rates, such as internet services, social networking sites, or streaming sites, have not only endorsed ML but have also actively influenced the scientific community to further develop the fields of Natural Language Processing (NLP) and computer vision. However, efficiently using these datasets offers several problems, causing this data to be commonly disregarded for science and clinical study. Obtaining excellent annotations and ground truth may also be difficult or impossible at this level. To overcome these issues, new computational approaches are required, and this thesis seeks to fill some of these gaps.

1.1.6. Popular Web-based Data Sources for Public Health Surveillance

Social media has emerged as a viable health communication platform [32,33]. Since there has been a huge growth in the number of people using social media to communicate information, researchers have been interested in analyzing social media activity for public health objectives. Furthermore, social media can cover subjects other than those addressed by traditional public-health data sources. Although some studies have questioned whether data from social media platforms would be useful for identifying outbreaks

[34,35], the study of social media content for health information has been a major topic [36,37]. To track and anticipate health incidents, [38] emphasizes the critical need to use social media platforms as a source of data while broadcasting pandemic alerts via their Twitter or Facebook accounts to quickly reach a larger public. As a result, social media posts and internet search behavior may be valuable sources of information concerning health epidemics. Some of the popular online data sources for PHS are:

1.1.6.1. Twitter

Twitter is a popular microblogging website where authorized members may publish tweets or retweet other' postings that unregistered users can read. With more than 300 million monthly active users, Twitter has become a reliable and timely source for determining the prevalence of health events in a community. H. Kwak, C. Lee, H. Park, et al. [39] undertook research to investigate Twitter's potential as a new source of information exchange. As social media postings from Twitter have become a dependable and rapid source for evaluating the occurrence of diseases in a community, effective and efficient methods for processing and examining health-related tweets must be created. For disease monitoring, variables such as location, volume, time [40], and public perceptions are typically taken into account. In a recent study [41], data from Twitter was used to identify information during various outbreaks that were valuable to various health agencies. C. Bosley et al. collected over 60 thousand tweets using a set of seven phrases, then examined and classified them, focusing on cardiac arrest and resuscitation [42]. Another study [43] proposed a model for predicting future influenza activity based on real-time Twitter data streams and historical datasets from the CDC. The data from Twitter was also used for investigated allergic activities [44]. The authors gathered tweets that referenced allergies as part of their research. In order to find Ebola-related tweets, a natural language processing technique [45] is used, focusing on four core subjects based on keyword clusters: risk factors, preventative education, disease patterns, and compassion. The vectors representing tweets were partitioned into clusters of similar words using an unsupervised approach [46]. Based on the similarity metrics of the clusters, the tweet is then categorized as disease-related or unrelated. N. El-bathy, C. Gloster, M. El-bathy, et al. [47] suggested a surveillance lifecycle architecture based on a unique genetic algorithm to make relevant data from a vast range of online data available more quickly and at a reduced cost. Twitter was also used to discover health issues such as respiratory, gastrointestinal, heat-related disease, and Influenza-Like Illness (ILI) symptoms circulated among people during huge gatherings [48]. Twitter data has been used to investigate various public health incidents, including allergy: [49], mosquito-borne disease: [50], dengue: [51,52], flu/influenza: [46,53–55], H1N1: [56–58] and other disease [59–62].

1.1.6.2. Instagram

Instagram is a photo-video sharing service created in 2010 and now has over 1000 million registered users. Health information shared on social media, particularly during a worldwide epidemic, has the potential to affect readers' or users' health habits and risk perception. Instagram, for example, is a free public sharing platform that may be used to provide important ideas or recommendations for remaining safe and healthy. In a study, the authors used Instagram to analyze the self-expression practices of chronic pain and concluded that Instagram is changing the way patients live with their diseases. In another research work, the authors [63] investigated Ebola-related social media posts on Instagram and Twitter, and their findings revealed that Instagram might be an optimum channel for communication and reaching the public during times of global health emergencies. Messages about Zika on Instagram have also been studied [64]. Instagram is a photo and video sharing network; thus, the image and video data available on the platform might be a useful source for disease surveillance.

1.1.6.3. Crowdsourcing

Crowdsourcing is a technique in which a vast, undefined group of volunteers or part-time workers participate in providing services, ideas, or material via a flexible open call [65]. Typically, a big group of people with diverse degrees of knowledge and experience work together to achieve a single objective. Google Consumer Surveys, Amazon Mechanical Turk, and private websites are among the most well-known crowdsourcing techniques. N. EO., K. SA., and B. JS [66] investigated the influence of Yelp.com reviews on foodservice on foodborne disease monitoring efforts and discovered that sickness reports collected online might supplement existing surveillance techniques. M. O. Lwin, S. Vijayakumar, and O. Noel [67] gathered crowdsourced information on mosquito bites, symptoms, and probable mosquito breeding areas and reported it to assist health authorities in early warning of dengue epidemics. A tool that combines crowdsourcing and text classification techniques was proposed to track misinformation about the Zika virus on Twitter [68,69].

1.1.6.4. Other Microblogs

A microblog represents a stream of text written by an individual comprising periodic and brief updates presented to the readers in reverse-chronological order. Compared with other social media, the information on microblogs is transferred in a truncated manner as the length of the post is limited. The short posts reduce users' time, and on average, a microblogger can post many updates in a day. Various popular microblog services other than Twitter also exist, such as Tumblr, Reddit (RD), Sina Weibo, Pinterest, etc. N. Yang, X. Cui, C. Hu C et al. [70] analyzed the content of a famous Chinese social medium (Sina micro-blog) and then used them to predict the flu outbreak in a region of

Beijing. To forecast seasonal flu, Z. Ertem, D. Raymond, and L.A. Meyers collected data from WordPress flu-related blogs (WordPress Flu), along with other online data sources such as Wikipedia and Twitter [71]. Other than Twitter, the majority of articles that obtained data from microblogs utilized Chinese microblogs, such as Sina Micro Index [72], Sina Weibo [73,74], and Tencent Weibo [75].

1.1.6.5. Internet Search Query

Internet Search Query analysis, particularly GT, is one of the most well-known sources of data for Internet-based monitoring in the field of healthcare. Scientists can get an overview of the current disease patterns by analyzing users' search queries as well as the location information recorded in their computers' Internet protocol addresses. One of the initial examples was Google Flu Trends, which started releasing real-time data publically in 2008. Based on flu-related terms searched by individuals on the Internet, it tracked flu epidemics throughout the world. It claimed to be one to two weeks quicker and almost as accurate as that of the CDC [76]. J.D. Sharpe, R.S. Hopkins, R.L. Cook, et al. examined the predictive performance of Google, Wikipedia, and Twitter-based surveillance. Consequently, Google Flu Trends outperformed in terms of sensitivity and positive predictive values [77]. Surprisingly, internet search behavior accurately predicted an increase in asthma-related emergency visits [78]. Aside from Twitter data, this article integrated internet users' Google search interests with environmental data to acquire information on asthma-related trips. Data from Google Trends has also been used to track infectious diseases such as tuberculosis [79] and influenza [80–82]. Some researchers [34,83,84] used web-based search queries to investigate and forecast the spread of an outbreak.

Some researchers have looked for data on other social media platforms. For example, S. Chaudhary and S. Naaz accessed digital health data for various diseases from the Practo, National Portal of India, and the Integrated Disease Surveillance Program (IDSP) Facebook pages [38]. Y. A. Strekalova analyzed the features of Facebook users who commented on CDC Facebook postings regarding the Ebola outbreak for more than seven months [85]. S. Gittelman, V. Lange, C. Crawford, et al. used Facebook "likes" to predict mortality, diseases, and lifestyle behaviors [86]. YouTube is another popular source for studying videos on various health issues, such as Ebola virus sickness [87]. Another study examined public responses to Zika virus-related YouTube videos to identify the video content [88]. According to S. Choi, J. Lee, M. Kang, et al. [89], short-text comments on news articles are popular in Korea and serve as a platform for people to express their own emotions and opinions. They examined the comments on news articles regarding the Middle East Respiratory Syndrome (MERS) outbreak to study public emotions.

1.1.7. Applications of Social Media-Based Surveillance Systems

Surveillance systems have been used in several modern applications in the domain of health informatics. These include disease prediction, misinformation tracking, worldwide awareness, etc.

1.1.7.1. Syndromic Surveillance-Based Disease Prediction

Syndromic surveillance has emerged as a promising method for predicting epidemics for public health objectives using data obtained from diverse sources prior to the accessibility of clinically validated data. The objective outcome is to decrease the spread in the population and to implement preventative actions. Social media data has been frequently utilized in recent years to estimate disease occurrences and detect disease outbreaks. Such data might be valuable for public health professionals in earlier detection of epidemics than the existing approaches. Typically, the information comes in the form of self-reported symptoms. According to research, monitoring systems in the healthcare domain can be used to forecast diseases that are a public health problem. Twitter data has been used to forecast syphilis [90], swine flu [91], tuberculosis [79], flu [92], Ebola [34], among other diseases. Another study presented the application of real-time Twitter data to complement the current disease surveillance efforts and examined disease occurrences of dengue and typhoid fever in the Philippines [51]. Similarly, another research [93] also employed surveillance systems using social media data to identify diseases.

1.1.7.2. Magnitude Estimation of Disease Over Time

Surveillance systems can be used to assess the magnitude of the epidemic [94]. Estimating disease levels in the future can help with planning, resource allocation, therapies, and prevention. The analysis offered by surveillance systems can be beneficial in determining the degree of the disease over time and making appropriate evaluations.

1.1.7.3. Event-Based Surveillance and Disease Prediction

Event-based surveillance is the rapid collection of data in an orderly manner concerning incidents that may pose a danger to public health. Data can come from various sources on the Internet, including media publications, online discussion platforms, routine reporting systems, personal information, and rumors. An event is defined in the context of a web forum as an excess number of news articles. The importance of the event is related to the number of postings about it. As a result, the event's influence on topic dispersion may be recognized based on the number of posts on the topic. When the number of posts on a topic surpasses the typical amount of postings, we may expect an event to occur at that time. The well-known instances of event-based surveillance systems include HealthMap, which records news connected to health occurrences, EpiSPIDER, etc. [95]. Over the last several decades, there has been a great increase

in the usage of social media-based public health intelligence monitoring approaches to give situational awareness of possible health concerns to assist surveillance actions [30,37]. N. Thapen, D. Simmie, C. Hankin, et al. [96] proposed DEFENDER, a software system with possible health event detection capabilities that monitors the Twitter stream and then reports the created events to users in a front-end User Interface.

1.1.7.4. Analyze User's Reactions to Health Events

Healthcare surveillance systems can be used to assess social media users' reactions to health promotion messages or events. T. Tran and K. Lee [97] extracted geotagged Ebola-related tweets from Twitter and collected 2 billion tweets in 90 languages between August and December 2014 to assess public reactions to Ebola. Another study gathered Twitter data to represent emotions at various phases of an upcoming health event [98]. S. Choi, J. Lee, M. Kang, et al. investigated the association between mass media and emotional public reactions during a nationwide MERS outbreak in Korea in 2015 [89]. E. K. Seltzer, E. Horst-Martz, M. Lu, et al. [64] argued that Instagram might be used to define public attitudes and also highlighted areas of public health concern. Another research highlighted Twitter users' interests in sharing Zika virus news, such as symptoms, accounts of Zika-infected pregnant women, and parental concerns [99]. Furthermore, R. Gaspar, S. Gorjo, B. Seibt, et al. demonstrated that, during the 2011 EHEC/Escherichia coli bacteria epidemic in Europe, amplification of public emotions via Twitter had a severe political and economic impact [100]. K. Liu, L. Li, T. Jiang, et al. examined differences in public reaction to various disease outbreaks by assessing search behavior [72]. In another study, I.C. Fung, K. Fu, Y. Ying, et al. used social media data to obtain insight into Chinese people's reactions to various epidemics [73].

1.1.7.5. Global Awareness of Events

Once the event has been identified, surveillance systems may be employed to track the general public's knowledge and impression of health-related occurrences. The sentiments expressed by users on social media platforms in response to an epidemic scenario represent their knowledge, attitudes, and perception. During epidemics, social media allows for sharing of public thoughts, ideas, and responses [101]. The high level of uncertainty around Zika may have prompted people to turn to social media to seek vital information [102], but it may also have led to the creation and propagation of conspiracy theories.

1.1.7.6. Tracking Misinformation

The ability of social media to offer information in real-time also allows for the rapid spread of disinformation among individuals during an ongoing pandemic that can have serious consequences.

However, few studies have been conducted to identify the propagation of disinformation regarding such events, and even fewer researchers have worked on widespread misinformation posted on social media platforms [103]. For example, Ghenai A and Mejova Y. employed machine learning techniques to detect Zika virus disinformation on the Twitter network [68]. Another research [104] investigated how to correct misconceptions regarding measles vaccination by observing the behavior of two social media groups. The authors later investigated the impact of clarifying the disinformation on users who responded to the measles vaccine. In the analysis of video content posted on YouTube during the Zika-virus outbreak, it was found that the dissemination of both facts and conspiracy theories occurred in a similar fashion [88], which was an eye-opening finding for health organizations. In order to combat the spread of such disinformation, greater attention must be paid to content uploaded on YouTube or any other online platform. Another study examined 1680 microblog entries and discovered that more than 20% of them included deceptive statements [75]. Indeed, the possibility of misinformation might be a barrier to using social media data in the domain of healthcare to fill the gap on how health organizations should receive a reaction and rectify misinformation [105].

1.1.7.7. Uncovering the Topics Popular During the Outbreak

Another use for health-related surveillance systems may be to extract the topics that were popular during the epidemic. The data gathered for monitoring purposes can be utilized to identify several concurrent occurrences during the outbreak [106]. Various issues, such as government/politics and the economy, were also recorded in order to study public reactions to the MERS epidemic in Korea in 2015 [89]. According to K. Rudra, A. Sharma, N. Ganguly, et al., during epidemics, views and sentiments are largely included in non-disease tweets (tweets that do not transmit any information regarding the epidemic condition) [41]. Similarly, data gathered for public health surveillance may be used to discover continuing issues during health incidents.

1.1.8. Limitations and Challenges of Social Media Based Surveillance Systems

This section highlights some of the limitations and challenges that surveillance systems encounter when employing social media data. While social media-based surveillance systems have resulted in advances in the early identification of epidemics and related events, several studies have questioned the outcome of these surveillance systems for the following reasons:

1.1.8.1. Noise

Noise is one of the difficulties encountered while gathering data. Data acquired from social networking sites may contain unrelated information to the goal [35,107]. Such information pertaining to disease

words has no relation to health. For example, postings including the keyword "Irish Flu" may create a huge number of flu-related activities. Unfortunately, the individual may post a status and be accused of being sick when they are not. In this way, false information can impact disease management for the public health department. To avoid such noise in data, fundamental text processing techniques such as feature weighting, tokenization, stemming, frequency-based algorithms, and so on are available [108]; nevertheless, further training is required to obtain the relevant data for further analysis.

1.1.8.2. Data Validation

Another problem with social media data is its validation. The usage of unauthorized data from social media may cause standardization, verification, and control issues [109]. T. Bodnar and M. Salathé concentrated their efforts on confirming the ground truth linked with a big volume of diverse social media data [110]. When developing prediction models, one of the most important components is the dataset. The outputs of the prediction models are heavily influenced by the dataset. Historical information, training data, and testing data are all included in the dataset for the prediction models. A substantial quantity of training data is necessary to train forecasting models, as well as testing data to validate predictions based on model training. Hence, data consistency is as important as data volume and quality of data from reliable sources [111].

1.1.8.3. Low Confidence

Low confidence is another issue that comes with social media-based statistics. According to one research [112], official websites are a more reliable source of vaccine information than social media. The quality of health-related data available online varies. Many social media data analysis tools may reveal spikes indicating that something notable is happening; however, this might be a reflection of panic rather than true instances of a disease epidemic. Additionally, users may post that they have the flu when they truly have a regular cold, or others may discuss the disease due to increased media hype. There is research that presents a conspiracy theory concerning the Zika virus epidemic on Reddit amid a public health crisis [113]. Such actions are the root causes of the lack of trust in internet data. In order to avoid this problem, further training would be necessary as a method of modifying the classification algorithms.

1.1.8.4. Demographic Bias

Although the location, username, number of followers, user profiles, and other information can be collected alongside tweet content, demographic information such as age, gender, and race are difficult to determine through tweets, making it difficult to determine who is posting about the outbreak and to whom public health efforts should be directed. Among the few studies that have looked at the users' profiles

[85], one looked at the Facebook comments in terms of sex and discovered that males wrote more posts per person than women. Another bias is that younger individuals are the most frequent social media users. Furthermore, the discovery that social media data is skewed towards active users, who are generally young adults, well-educated, and have a decent income, lends credence to the bias [114]. This study employed machine learning techniques to demonstrate that sociodemographic features have an important influence in reporting disease-related messages on social media. As a result, social media users are thought not to reflect the entire population [115]. Any conclusion based on social media data excludes persons who do not use such platforms, who are likely to be the most vulnerable in the population, and who frequently do not wish to discuss their health experiences publicly. It is also possible that people who are in pain, unwell, aged, or incapacitated will be less likely to be active users.

Another bias caused by a population's demographics is language. Despite the fact that the heart of the epidemic/event is in a nation where English is not an official language, most research focus on a single language (English) [116]. Very few studies have shown research that relies on languages other than English to deal with disease tracking. For example, M. U. Ilyas has regionally filtered disease terms in both Arabic and English tweets [117]. Furthermore, the author discovered that tweets in English made up a relatively small percentage of the total. Their method was validated by a significant degree of association between the actual occurrence of MERS-Coronavirus infections and disease-related tweets in Arabic. Another research [118] discovered a variation in the responses of persons from various linguo-cultural backgrounds to the same epidemic.

1.1.8.5. Privacy Issues

There is also debate concerning ethical problems when getting data from social media [119], such as the privacy of information gathered via social media for health purposes [120]. Despite the fact that social media data is widely available, individuals may object to their postings or data being used for research purposes. Users' confidentiality concerns, such as diagnoses based on publicly available data or concerns over algorithms that reveal undeclared personal features, must be taken into account by social media platform hosts. Only a few studies have shown interest so far in addressing the privacy concerns while accessing the web-data [119,121]. As a result, there is a larger possibility of taking privacy into account while researching social media data.

1.1.8.6. Lexical and Linguistic Variability

Though communication via social media aids in extracting healthcare information, it is difficult in terms of semantic understanding of the language. Because social media writings are informal and vague, simply matching keywords may not capture their semantic interpretation, resulting in an incomplete result [122].

Machine learning techniques are increasingly being used in healthcare, allowing for more effective estimation and prediction of health outcomes from vast administrative data sets. In the following sections, we explained Machine Learning techniques and discussed how we build upon them toward more accurate pandemic time-series prediction.

1.2. Machine Learning

Since their origin, humans have employed a range of tools to do various tasks in a more effective manner. The inventiveness of the human brain led to the construction of various technologies. These technologies make people's life simpler by addressing a wide range of demands, including computing, industry, and travel. And, Machine learning is one of them.

The automatic recognition of significant patterns in data is referred to as Machine Learning [123]. These patterns may then be utilized to either improve our understanding of the existing environment (for example, identifying risk factors for disease) or to forecast the future (for example, predicting the infected cases). It has become a standard technique in practically every endeavor that demands extracting information from massive data sets during the last few decades. We are surrounded by ML-based technologies such as search engines learning to deliver the best results, filtering email messages by anti-spam software, and safeguarding credit card transactions by software learning to identify fraud. Intelligent personal assistance software on smartphones learns to understand voice commands, while digital cameras learn to distinguish faces. Accident prevention systems are integrated into automobiles utilizing ML algorithms. In addition, ML is frequently employed in scientific applications such as astronomy, medicine, and bioinformatics [124].

ML is a technique for training machines to process data more effectively. With so many datasets accessible, the need for ML is increasing to analyze the extracted information from the collected data. ML is used in many sectors to extract essential data. Many experiments have been carried out in order to identify how to train machines to learn without being explicitly programmed.

1.2.1. When Do We Need Machine Learning?

When is it necessary to use ML instead of explicitly instructing our computers to do a task? Two features of a particular situation that may necessitate the deployment of programs that learn and develop depending on their experience are the complexity of the problem and the necessity for adaptability [124].

- **Problems Which Are Too Complex to Program**

Human-performed tasks: There are countless tasks that we humans undertake on a daily basis, but our insight into how we accomplish them is insufficient to implement a well-defined program—for example, speech recognition, visual comprehension, driving, etc. State-of-the-art ML algorithms, which learn from their experience, get pretty good outcomes in all of these tasks when dealing with a significant number of training examples.

Tasks beyond Human Capabilities: ML techniques are popularly used to analyze huge and complex data sets, such as e-commerce, Web search engines, analysis of genetic data, weather prediction, translating clinical records into knowledge, and astronomical data. With the availability of more digital data, it becomes evident that valuable information is hidden in data records that are much larger and more complicated for humans to make sense of. Finding meaningful patterns in huge and complicated data sets is an interesting issue. Combining learning algorithms with computers' practically endless memory capacity and continually improving processing power opens up new possibilities.

- **Adaptivity**

The rigidity of programmed tools is one of its limitations; once the program is developed and implemented, it remains unaltered. On the other hand, several tasks vary from one user to another or over time. ML programs provide a solution to such problems since, by definition, they are adaptive to changes in the environment with which they engage. Examples of successful machine learning applications to such problems include programs to decode handwritten text, spam detection, speech recognition programs, etc. [125,126].

1.2.2.Types of Machine Learning

ML incorporates a number of techniques to address data issues (Figure 1.2). Data scientists indicate that there really is no one-size-fits-all solution that is suitable for resolving a problem. The approach employed is determined by the type of problem to be addressed, the number of variables, the appropriate model to utilize, and other factors.

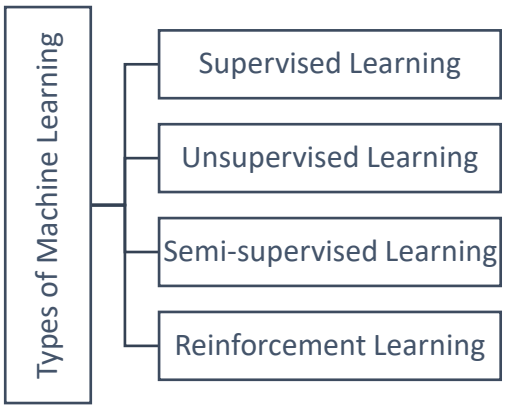


Figure 1.2 Types of Machine Learning Algorithms

1.2.2.1. Supervised Learning

Supervised learning is the ML approach where machines are trained on labeled input data to yield the desired output. The input dataset is divided into two categories: train and test. Each labeled training data set comprises an input value and a desired target output value. An output variable must be forecasted or classified in the training dataset. All algorithms detect patterns in training data and implement them to test data to make predictions or classify it. Supervised learning identifies patterns that relate variables to measured outcomes and maximizes accuracy when predicting those outcomes. For example, an automatically fitted regression model is a supervised learning technique [127].

1.2.2.2. Unsupervised Learning

Unsupervised learning refers to the ML technique in which models analyze and learn patterns from unlabeled data. When new data is presented, it uses previously learned features to determine the class of the data. Feature reduction and clustering are its principal applications. Unsupervised learning uses the natural qualities of the input data set to recognize trends and patterns without explicitly designating one column as the desired output [128]. Unsupervised methods include principal component analysis, revealing underlying covariance patterns in observed data [127].

1.2.2.3. Semi-Supervised Learning

Semi-supervised learning combines supervised and unsupervised ML approaches [129]. This combination often comprises a small quantity of labeled data and a significant amount of unlabeled data. Initially,

related data is grouped using an unsupervised learning method, and it then assists in labeling the unlabeled data. Semi-supervised learning, a type of hybrid, is utilized in situations when prediction is desired, but the bulk of data points lack outcome information.

1.2.2.4. Reinforcement Learning

Reinforcement learning can be traced back to the initial days of cybernetics and has since been applied to statistics, psychology, neuroscience, and computer science research. It has sparked a surge of interest in the machine learning and Artificial Intelligence (AI) sectors over the last five to ten years. Several real-time machine learning systems learn using a penalty-reward system based on feedback on its predictions rather than by fitting a model to a previously acquired data set [130]. This method, known as reinforcement learning, has the potential to be employed in online data collecting tools and monitoring. Reinforcement learning is an ML training method in which an agent learns to take actions in an environment to maximize reward. In this learning, a computer program grants access to a dynamic environment to achieve a specified objective. The program receives feedback in the form of rewards and penalties as it navigates its drawbacks [130].

1.2.3. Machine Learning Applications in Healthcare

The goal of machine learning in computer science is to improve the machine's efficiency and reliability. Machine learning applications are everywhere and are used in many real-world applications. It is essential in several areas, such as healthcare and medical data protection. In healthcare, the machine serves as an extension of the doctor's brain and a force multiplier. A machine's job is thus not to replace the doctor but instead to assist them in providing better service and care. Some of the applications of machine learning in the healthcare industry are:

1.2.3.1. Identifying and Diagnosing Diseases

ML-based approaches aid in the more accurate and speedy investigation of medical data, allowing clinicians to be more precise with accurate and prompt diagnoses. ML can detect patterns of certain diseases in patient electronic health information and alert clinicians to any irregularities [131].

1.2.3.2. Drug Discovery and Manufacturing

Machine learning is being used in the early phases of drug development, which is among the most efficient medical applications. This includes research and development technologies like precision medicine and next-generation sequencing, which can help in the identification of alternative therapy approaches for complicated conditions. Currently, ML approaches use unsupervised learning to detect

patterns in data without generating predictions. Microsoft's Project Hanover [132] utilizes ML-based technologies for a variety of different purposes, including the creation of Artificial Intelligence-based technology for cancer therapy and the customization of Acute Myeloid Leukemia medicine combinations

1.2.3.3. Medical Imaging Diagnosis

Computer vision is a ground-breaking technology that is made possible by both ML and Deep Learning (DL). Machine learning methods used in computer-aided detection and diagnosis can assist clinicians in interpreting medical imaging data and decrease interpretation times [123]. As machine learning is becoming more accessible and their analytical capacity is developing, more data sources from numerous health imaging are expected to become a part of this AI-driven diagnostics.

1.2.3.4. Personalized Medicine

Due to the exceptional performance of ML models when used with complicated large data, the advent of ML applications over the last decade has resulted in considerable gains toward the deployment of personalized medicine techniques for enhanced health care. Personalized therapies are not only more successful when individual health is combined with predictive analytics, but they are also ideal for additional research and improved disease evaluation. Clinicians are now constrained to choose from a limited number of diagnoses or estimate the patient's risk based on his clinical history and genetic information. However, ML is making enormous breakthroughs in medicine. One example is IBM Watson’s Oncology which utilizes individual medical data to help in developing multiple therapeutic options [133]. Additional gadgets and biosensors with improved health measuring functionalities will enter the market, allowing more data to be readily available for these kinds of medical systems.

1.2.3.5. Behavioral Modification

Behavioral modification is an important component of preventative medicine, and as ML has become more prevalent in the health sector, a spate of start-ups has emerged for cancer identification and control, patient treatment, and many more. To address mental health issues, supervised machine learning approaches can be employed. Using DL techniques like Convolution Neural Network (CNN), Recurrent Neural Network (RNN), and Artificial Neural Network (ANN), a sophisticated model can be created that can predict a person's mental health based on his expressions, activities, and bodily movements [134].

1.2.3.6. Maintaining Health Records

Maintaining health records is a time-consuming process. Though technology has assisted in inputting data, most of the processes usually take considerable time to finish. ML's primary job in the health sector

is to simplify operations to save effort, time, and cost. Ciox, a healthcare technology firm, for example, employs ML technologies to improve healthcare information administration and exchange [135]. The objective is to increase access to medical data, streamline the company's operations, and enhance the accuracy of healthcare information. It also created smart charts, which are used to recognize and collect medical data from multiple health records in order to combine a patient history into a single digital profile.

1.2.3.7. Clinical Trial and Research

Clinical studies are time and money-consuming and might take years to complete in many cases. ML technology can aid in predicting clinical trial outcomes, resulting in shorter drug approval timeframes, reduced costs, and more funding for the development of novel treatments [133]. ML has also been utilized to enable real-time monitoring for trial participants, as well as to establish the proper sample size to be evaluated and to make use of the capabilities of electronic records to avoid data-based errors.

1.2.3.8. Crowdsourcing

Crowdsourcing allows researchers and practitioners to examine vast amounts of data produced by people who have given their approval. This real-time health data will have a big impact on how healthcare is perceived in the future. For example, users may use Apple's Research Kit to access interactive programs that employ machine learning-based face recognition to cure Asperger's and Parkinson's diseases [133]. With IoT advancements, the healthcare industry is always looking for new ways to leverage the data to manage challenging situations and improve treatment and diagnosis.

1.2.3.9. Better Radiotherapy

Radiology is among the most common use cases of ML in the health sector. In clinical image analysis, several distinct variables might arise at any time. Many diseases such as tumors, cancer foci, and others are challenging to depict mathematically. Because ML-based methods learn from varied samples, identifying and detecting factors becomes easier. Machine learning can aid in the automation of both the diagnosis of cancer and the physiological structures in healthy organs and the accurate selection of the optimum radiation dose [136].

1.2.3.10. Outbreak Prediction

Machine learning and AI-based technologies are currently being utilized to monitor and predict outbreaks [137]. Intelligent Models using ML for disease Prediction, whether constructing a model to assist doctors or even avoiding disease transmission in a certain area, are increasing day by day. Researchers now have

access to vast volumes of data acquired from various sources, including real-time social media and online data. Clearly, the disease outbreak parameters provided are enough for ML decision support algorithms to anticipate an epidemic properly.

1.3. Public Health Surveillance Systems and Machine Learning

Public health surveillance systems analyze patterns in environmental circumstances, disease incidence, and health behaviors to assign resources to sustain healthy communities. The "curse of dimensionality" related to large data has been partially relieved using ML models within the data science field, especially in scenarios where forecast or hypothesis formation is the analytic goal instead of hypothesis testing [127]. Traditional epidemiology may be supplemented by machine learning to assist public health in sorting through vast volumes of data. ML may improve predictive modeling, cluster analysis, and social network studies, all of which can improve the sensitivity of surveillance systems. Machine learning is used as one method in epidemiology and public health monitoring to do causal inference analysis, diagnostic and prognosis research, genome-wide association studies, geospatial applications, and forecasting. Such machine learning algorithms have the potential to improve aberration detection as well. The overview of ML workflow is depicted in Figure 1.3.

In recent years, machine learning has garnered a lot of attention, particularly when it comes to finding patterns in images or raw data. M. Conway, D. O'Connor [28] reviewed the conjunction of natural language processing and machine learning with social media platforms to support the analysis of massive datasets for population-level mental health research. Among different methodological variations of machine learning, some architecture stands out in popularity. For instance, we noted that the k-Nearest Neighbor (k-NN) classifier's precision was superior to several other Machine Learning classifiers such as Multinomial Naive Bayes (NBM) Modal, Naive Bayes (NB), and Support Vector Machine (SVM) [49], for classifying the tweets between two classes-i.e., real occurrences of allergy or awareness tweets. Similarly, K. Lee, A. Agrawal, A. Choudhary, et al. [44] showed that the best text classification performance was obtained using Multinomial Naive Bayes Modal when compared against the other classifiers such as NB, Random Forest (RF), and SVM. The author [43] exhibited a model using the multilayer perceptron with a backpropagation algorithm on Twitter data to predict the weekly status of the United States population infected with ILI. Even several supervised machine learning algorithms were studied for detecting the personal health experience tweets [107] and used the deep gramulator approach to improve precision when applied to independent test sets.

Unlike supervised algorithms, unsupervised classification algorithms do not require labeled data sets to predict the outcome. Consequently, unsupervised classification methods appear to be a more tempting alternative in the text analysis process; nonetheless, they may be more challenging to reach equivalent

accuracy than supervised ones. When L. Sousa, R. de Mello, D. Cedrim, et al. categorized tweets using supervised and unsupervised algorithms, they discovered the same phenomenon [50]. They concluded that topic modeling, such as Latent Dirichlet allocation (LDA) [138], one of the unsupervised techniques, gives less control over topic material than a standard classifier, particularly on a naturally noisy media channel. As a consequence, Multinomial Nave Bayes, a supervised classification approach, was investigated for classifying Twitter content.

Although machine learning is commonly used in data science, it has been adopted by certain public health experts, and professionals have adopted it as well. For example, ML has been employed for geographical and spatiotemporal monitoring [137], epidemic detection and monitoring [139], and identifying patient features related to health outcomes [140]. Variants of existing ML algorithms have been used to estimate rehospitalizations [141], disease transmission [142], and identify suicidality among individuals [143], construct an early warning system for adverse drug reactions based on internet data [144], among other uses.

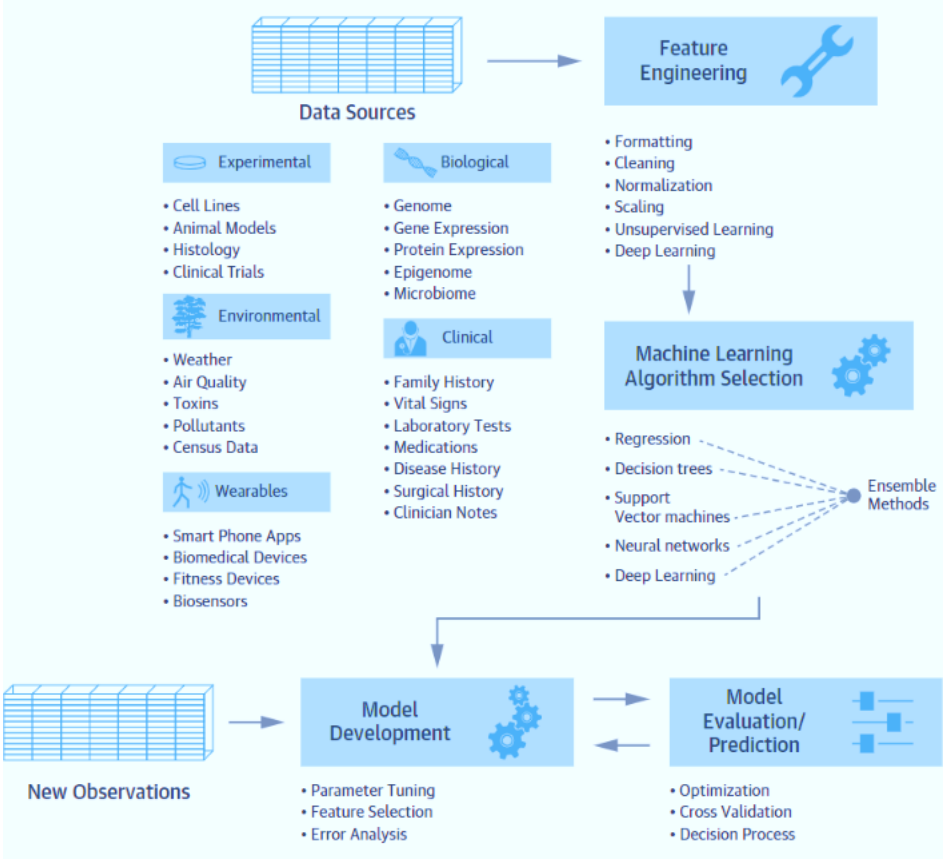


Figure 1.3 The Overview of Machine Learning Workflow [145]

1.4. Motivation of Study

The Public Health Surveillance aims to offer a factual foundation for authorities to define priorities, organize programs, and take measures to promote and protect the public's health.

For the PHS, it is important to identify and select a health problem for surveillance. The world's populations face a slew of health issues. Certain issues pose an urgent threat to health, while others are long-term issues with relatively consistent incidence and prevalence among those affected. Influenza outbreaks are a prime example of the former; atherosclerotic cardiovascular disease and colon cancer are examples of the latter. Health issues change for different people and locations, and what is an acute hazard in one population may be a chronic concern in another. A malaria outbreak in the United States, for example, would be an urgent concern in 2006, but malaria in Africa is a persistent problem.

Because performing surveillance for a health problem requires time and resources, exercising caution when picking health concerns for monitoring is necessary. In some countries, the selection is based on disease prioritization criteria, a review of existing mortality and morbidity data, an understanding of diseases and their spatial and temporal trends, and sentiments on public issues. The following criteria have been devised for identifying and prioritizing health concerns for surveillance [146–149]:

- Public health significance of the problem, which includes incidence, prevalence, severity, sequela, disabilities, mortality caused by the problem, socio-economic impact, communicability, outbreak potential, public perception, and concern and international requirements
- Ability to prevent, control, or treat a health problem
- Capacity of the health system to execute control measures for the health problem: reaction time, economics, resource availability, and what surveillance of this event is required.

With the significance of machine learning, actionable insights and knowledge can be obtained from the massive online datasets for influencing decisions and enhancing people's lives. Health data is the best contender for directly changing people's lives. The manner in which we process this data has the potential to revolutionize our civilization.

Prevention and control of disease need knowledge of the major medical causes of disease as well as exposures or risk factors. In 2017, at the global level, the top 10 causes of maximum deaths were related to disease, illness, and other public health factors (Figure 1.4). It was noticed that non-communicable diseases dominated worldwide death numbers, with cardiovascular diseases being the leading cause of death globally. The second biggest cause was cancer. Even in the same year, dominating causes of death in India were diseases and other health factors (Figure 1.5). However, ML has made great advances in the healthcare domain, yet to reduce the burden of diseases, more promising systems are needed for making accurate and timely decisions.

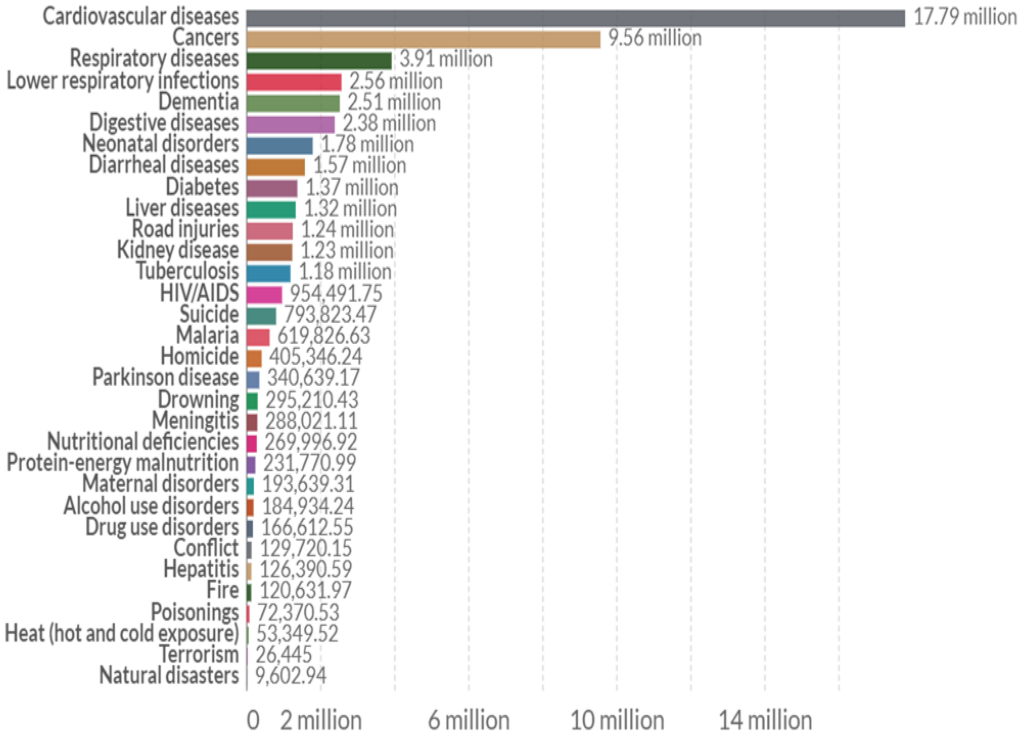


Figure 1.4 Number of Deaths by Cause, World, 2017 [150]

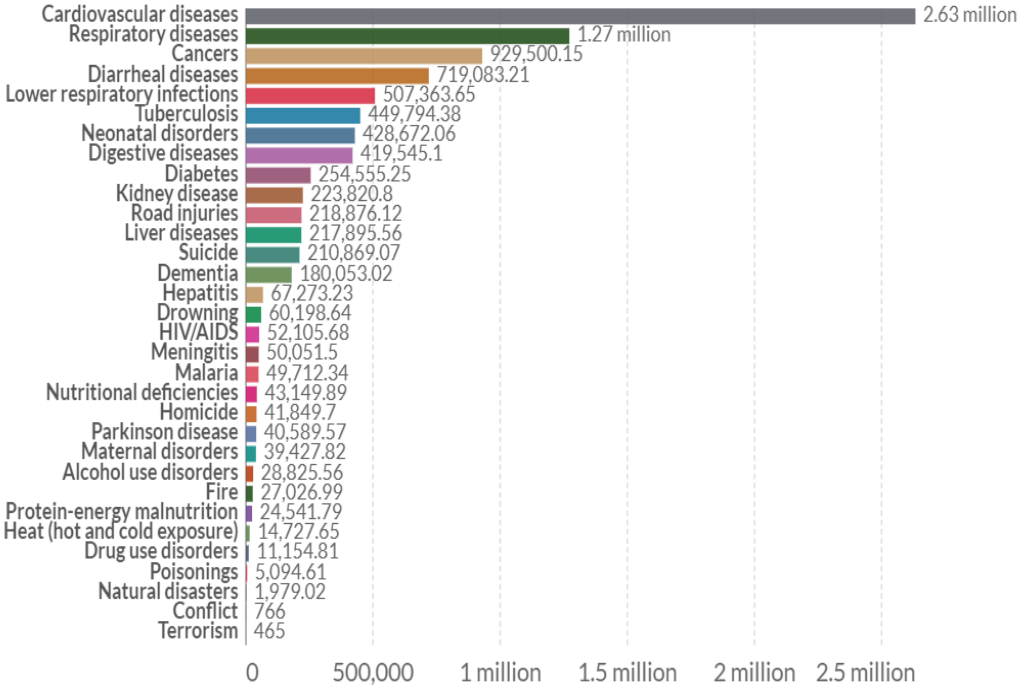
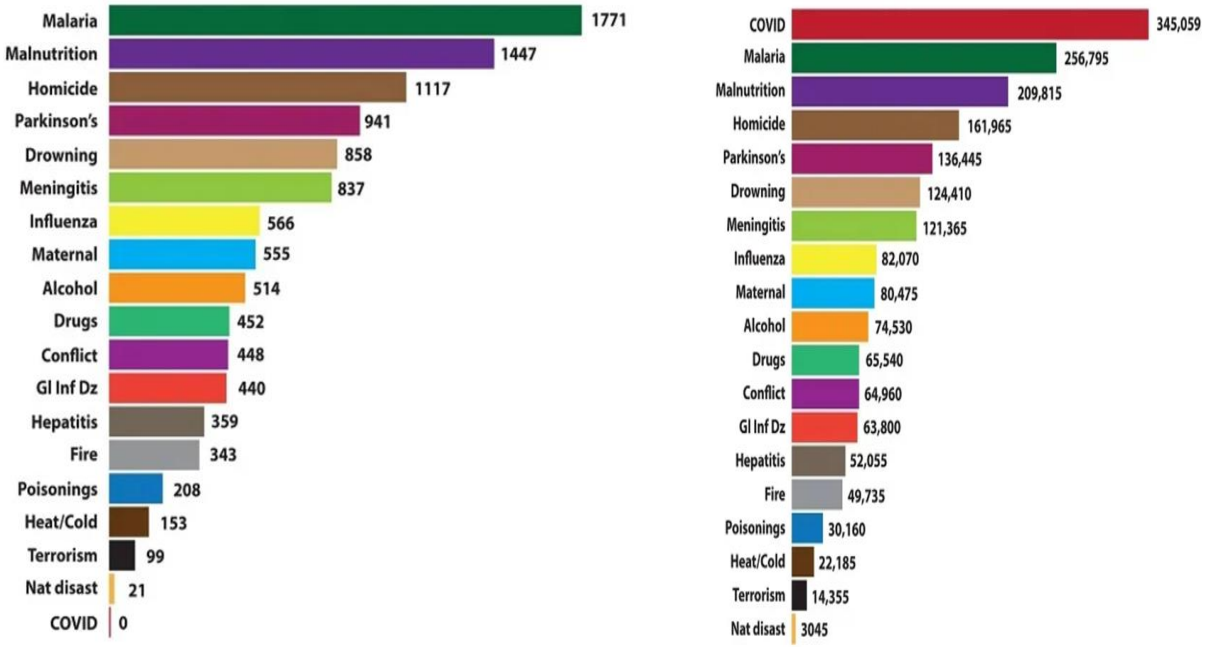


Figure 1.5 Number of Deaths by Cause, India, 2017 [150]

In 2020, the pandemic of COVID-19 was a wake-up call for the public health realm, which has become the leading cause of death worldwide and a top global public health priority (Figure 1.6). Ever since the onset of COVID-19, a total of 214 countries have reported cases of infection and deaths due to it (Figure 1.7). The COVID-19 pandemic has made health practitioners more aware of the need for technologies for more representative research.



(a) (b)

Figure 1.6 Leading Causes of Deaths Worldwide (a) 01/01/2020 (b) 24/05/2020 [151]

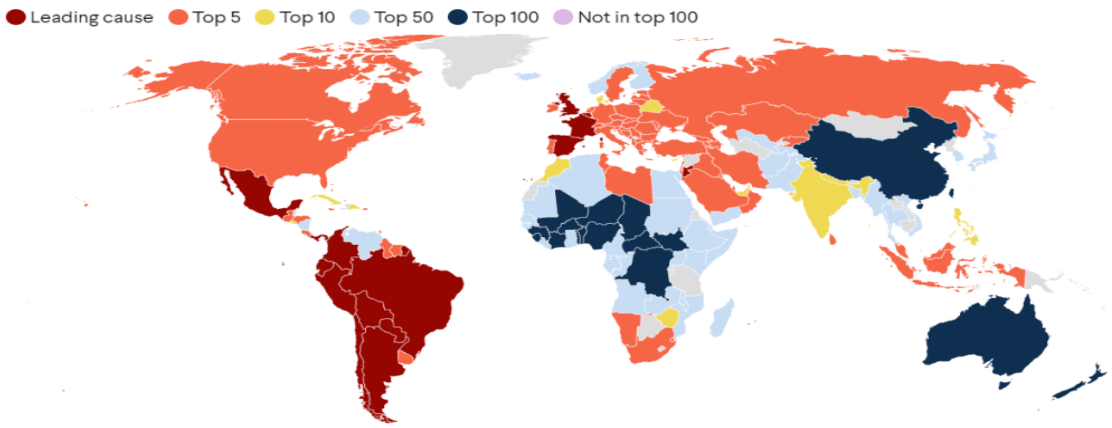


Figure 1.7 Ranking COVID-19 Deaths Among All Causes of Death Since Start of 2020 [152]

Even in the year 2021, COVID-19 remains one of the leading causes of death in India (Figure 1.8). Health research systems were already under strain when COVID-19 brought significant challenges, as well as making the existing challenges in health research systems more visible and significant. Regarding this, we directed our research work towards the recently ongoing COVID-19 pandemic to help academicians and researchers.

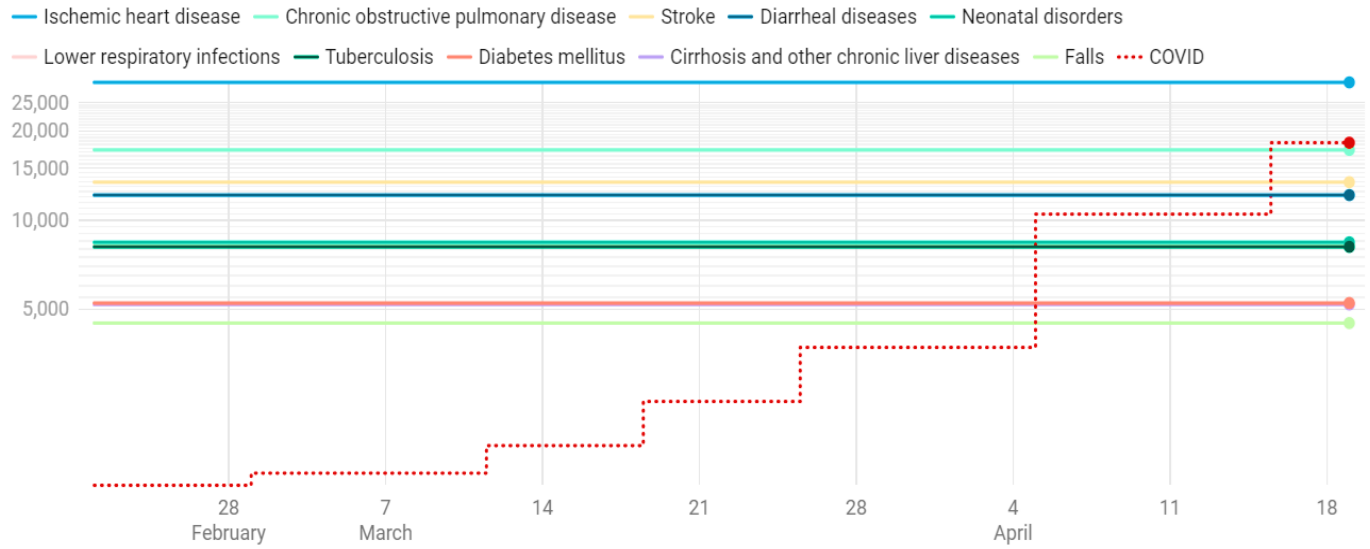


Figure 1.8 Leading Causes of Death, India,2021 [153]

1.5. Research Objectives

The primary aim of this research is to try to overcome restrictions and build novel and innovative methods for inspecting technology-driven techniques. In order to accomplish this aim, the following Research Questions (RQs) have been established:

- **RQ1**: Which machine learning techniques are commonly used to develop social media-based surveillance systems in the health sector?
- **RQ2**: How might machine learning techniques be used to enhance pandemic prediction?
- **RQ3**: What significance does feature extraction play in improving pandemic prediction?
- **RQ4**: What is the impact of combining pandemic data with information like environmental risk factors, meteorological conditions, demographic data, and so on?
- **RQ5**: Is it possible to enhance prediction performance by combining data from multiple online platforms?

For finding the solution to the above sub-queries, the following Research Objectives (ROs) are finalized:
Research Objective 1: To research and implement suitable state-of-the-art Machine Learning or Deep Learning algorithms that could improve the predictive performance of public health surveillance systems.

Research Objective 2: Considering the conjunction of data with information like environmental risk factors, weather conditions, demographic information, etc.

Research Objective 3: Propose a suitable feature extraction algorithm while addressing the significance of varying weights to different categories.

Research Objective 4: Exploring the applicability of using multiple social media platforms simultaneously to get accurate and timely predictions of epidemics.

The detailed description of the identified research objectives is as follows:

Research Objective 1: In this objective, machine learning and deep learning techniques are implemented for infectious disease prediction. In this objective, an epidemic compartmental model is proposed, which implements the evolutionary algorithm, Long Short-term Memory (LSTM) neural network, and Latent Dirichlet allocation based, Pandemic-LDA (PAN-LDA) module for predicting the infectious disease.

Research Objective 2: As health decision-makers require a variety of information, including health determinants (social-economic, environmental, behavioral, and demographic factors) and the context-specific environments in which the healthcare system works, therefore in this objective, we examined the impact of related multi-source data on the COVID-19 pandemic. In addition, a short-term prediction based on the effect of health determinant variables was proposed to forecast COVID-19 cases.

Research Objective 3: The feature extraction process is extremely crucial for the effectiveness of Machine learning applications. Its primary purpose is to extract relevant information from input data in a compact manner while reducing noise and redundancy to improve the accuracy of ML models. Therefore, in this objective, we have proposed two feature extraction algorithms. One algorithm used semantically and morphologically similar word embedding clusters as features to improve clustering performance. The other algorithm made use of the historical cases and the corresponding news articles to provide better features to machine learning algorithms and improve prediction accuracy.

Research Objective 4: Overall, there are relatively few studies in the present literature that use big-data availability from multiple social media and online platforms, such as Facebook, Twitter, Reddit, Google News (GN), and others, to model and predict the number of confirmed cases during an epidemic. To address this gap, we have proposed an epidemic model using data from multiple online platforms to improve prediction performance.

In this context, for fulfilling the requirement of RQs and ROs, Table 1.1 demonstrates the mapping among RQs, ROs, and publications.

Table 1.1 Aligning of Research Questions, Research Objectives, and Publications

ROs	RQs	Publication(s)
RO1	**RQ1, RQ2**	Aakansha Gupta, Rahul Katarya. "A Deep-SIQRV Epidemic Model for COVID-19 in India to Access the Impact of Prevention and Control Measures".**[Communicated]** Aakansha Gupta, Rahul Katarya. (2021). "PAN-LDA: A latent Dirichlet allocation based novel feature extraction model for COVID-19 data using machine learning". **(Published, SCIE, Impact Factor: 4.589).**
RO2	**RQ3**	Aakansha Gupta, Rahul Katarya. "Spatial and temporal distribution of the COVID-19 growth rate in India and its correlation with influencing factors". **[Communicated]** Aakansha Gupta, Rahul Katarya. (2022). "Possibility of the COVID-19 Third Wave in India: Mapping from Second Wave to Third Wave". **(Published, SCIE, Impact Factor: 1.778).** Aakansha Gupta, Rahul Katarya. "Human mobility based Pandemic Prediction Model" 3rd IEEE International Conference on Advances in Computing, Communication Control and Networking (ICAC3N–21). **IEEE**
RO3	**RQ4**	Aakansha Gupta, Rahul Katarya. (2021). "PAN-LDA: A latent Dirichlet allocation based novel feature extraction model for COVID-19 data using machine learning". **(Published, SCIE, Impact Factor: 4.589).** Aakansha Gupta, Rahul Katarya. (2022) "Deep embedding for mental health content on social media using vector space model with feature clusters". **(Published, SCIE, Impact Factor: 1.536).** Aakansha Gupta, Rahul Katarya. "Analyzing the Effects of Text Representations on the Performance of Document Clustering in Public Health Tweets" 4th International Conference on

		Computational Intelligence & Data Engineering (ICCIDE-2021).**Springer** Aakansha Gupta, Rahul Katarya. "A Novel LDA-based Framework to Forecast COVID-19 Trends". 4th International Conference on Innovative Computing and Communication (ICICC-2021). **Springer** Aakansha Gupta, Rahul Katarya. "Improving document representation using KPCA and clustered word embeddings" 2021 Fifth International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT).**IEEE**
		Aakansha Gupta, Rahul Katarya. (2020). "Social media based surveillance systems for healthcare using machine learning: A systematic review". **(Published, SCIE, Impact Factor: 6.317)**
RO4	**RQ1, RQ5**	Aakansha Gupta, Rahul Katarya. "A Deep-SIQRV Epidemic Model for COVID-19 in India to Access the Impact of Prevention and Control Measures".**[Communicated]** Aakansha Gupta, Rahul Katarya. "COVID-19 cases prediction based on LSTM and SIR model using social media" 7th International Conference on Data Science and Engineering (ICDSE 2021).**Springer**

1.6. Outline of the Thesis

The thesis consists of six chapters describing the entire study in a very concise and precise way. Each chapter is summarised below:

Chapter 2: In this chapter, we present a brief description of the state-of-the-art methods and previously developed methods and their merits and demerits to discuss the social media-based public health surveillance system using machine learning approaches. During this research, multiple approaches are studied used for public health surveillance systems using machine learning techniques. This chapter also covers the background detail of machine learning methods for public health surveillance.

Chapter 3: In this chapter, a machine learning and DL-based epidemic model, called deep-SIQRV, is proposed, which considers the influence of prevention and control strategies to simulate the spread of COVID-19 and forecast infectious disease cases. Also, the framework of the proposed epidemic model is discussed in detail. In order to enhance the prediction performance, an LDA-based PAN-LDA model has been used to incorporate the textual features of the media reporting and the public awareness of the pandemic from the various online platforms. In order to validate the authenticity of the proposed method,

our experimental outcome is compared with the other standard compartment models that illustrate the supremacy of our proposed model.

Chapter 4: As the dissemination of the pandemic is determined by several factors such as socio-economic, environmental, behavioral, demographic factors, etc. Therefore, in this chapter, we have analyzed the spatial and temporal trends in the COVID-19 monthly growth rate, the factors associated with the disease distribution, how these factors affected the growth rate of the disease, and the interaction effects of the different variables on the pandemic growth rate. Later, we proposed a short-term fixed-effect multiple regression model to predict the confirmed cases during the initial stages of the coronavirus second wave. We also analyzed the possibility of the COVID-19 third wave based on the factors related to the spread of the epidemic during the initial phases of the coronavirus second wave.

Chapter 5: In order to improve text clustering, this chapter proposed a feature extraction algorithm that uses semantically and morphologically similar word embedding clusters based on the vector space model. The proposed method can reduce the feature dimensions compared to conventional VSM-based models. Despite the simplicity of the proposed method, the experimental outcomes illustrate the supremacy of our proposed approach by enhancing the clustering performance compared to other text representation methods.

Chapter 6: In this chapter, a feature extraction algorithm called PAN-LDA, a modification of LDA, is proposed that uses the historical cases of the coronavirus and the corresponding news articles in the common LDA to obtain a new set of features. The generated features are introduced as additional features to Machine Learning algorithms to improve the forecasting of time series data.

Chapter 7: The importance of online and social media platforms is vast, deep, and multi-directional. So in this chapter, the significance of online and social media platforms is explored to improve the prediction performance of the COVID-19 pandemic. An epidemic model is proposed that incorporates the features from data collected from multiple online sources such as Twitter, Reddit, and Google news to improve the prediction performance. We compare the results with existing models and show that the language information processing of relevant posts can help enhance the prediction accuracy of the epidemic model.

Chapter 8: In this chapter, the conclusion, future scope of the current research, and limitations have been discussed. Currently, machine learning-based models produce significant results for public health surveillance systems. In the future, ML-based models can be used to encode social media data in

epidemic models. Finally, a few futuristic suggestions and plans have been suggested to enhance public health surveillance. The future of online information and machine learning techniques is very bright and strongly opens new flaps for the world.

1.7. Chapter summary

This chapter covers the overview of public health surveillance along with the machine learning techniques. We also discussed the impact of online data on enhancing the public health surveillance system performance. In this chapter, we discussed the content and description of each chapter. Each chapter includes some unique concepts and ideas to support the title and objective of the thesis strongly. The chapter briefly explains public health surveillance, machine learning, and embedding of social media data for public health surveillance. The objectives and scope of the entire work with motivation to study are also described in detail.

Chapter 2 METHODICAL LITERATURE REVIEW

Recent advances in the production, collection, and curation of data have resulted in a new and complex set of resources available to guide public health decisions. At the same time, improvements in the accessibility of advanced statistical and computational tools have expanded the capacity of public health professionals to conduct rigorous empirical analyses and develop decision support tools that are grounded in observed data. In this chapter, related work comes from three different research fields: Public health surveillance, where we find the inspiration for this thesis; ML & Statistics, where we find the methods and techniques to address the defined problems; and online sources of data, where we find the information insights that help to design better computational models. Section 2.1 covers the overview of the chapter. Section 2.2 explores the entire review process: the study plan selection of studies and data set extraction. Section 2.3 explains the considerable description of the studies that used machine learning algorithms for public health surveillance. Section 2.4 describes the research gaps and limitations of the previous studies. Section 2.5 concludes with the summary of the chapter.

2.1. Overview

Real-time surveillance in healthcare informatics has evolved as a growing area of interest among researchers worldwide. The advancement of this discipline has aided in the implementation of different projects linked to public health informatics. Surveillance systems in the field of health informatics that use social media data to anticipate disease outbreaks and monitor diseases have been developed. In recent years, the growth of social media platforms, especially Twitter, has enabled real-time syndromic monitoring, allowing for quick analysis and input to those in charge of follow-ups and investigating suspected outbreaks.

Syndromic surveillance systems collect data in order to give a fundamental scenario for all contagious infectious diseases. These systems may vary depending on the data source, their planned duration, and how data is recorded and acquired. These systems may use traditional data and real-time data from various social media platforms. In the field of healthcare, these systems usually focus on the early identification of illness clusters and the symptom period before the confirmation of a particular disease by any clinical unit or laboratory and to mobilize the rapid response. Surveillance systems are usually concerned with the systematic collection, analysis, and interpretation of the collected data along with the detection, confirmation, and reporting of disease and also considering the public health response. Traditional biosurveillance relies on clinical encounters to collect information, which is a time-consuming process. Also, traditional pandemic surveillance is mostly a manual process that causes a delay of one to

two weeks in the availability of the data by clinical diagnosis [154]. In the last few years, the availability of web-based data sources emerged as an extension to traditional surveillance systems [23] and has sustainably contributed to infectious disease surveillance by providing real-time statistics and reducing the cost of public health [24]. It can be noted how rapidly events can be detected in real-time using internet-based surveillance data when an ordinary individual's social media post has led to a tremendous increase in public engagement with skin cancer prevention [25]. Despite the other uses of social media [155], the role of monitoring social media can be explored in making healthcare decisions [156]. Also, the early detection of diseases and immediate public health response has led to the need for new approaches and technologies to reinforce the capacity of traditional syndromic surveillance systems.

When paired with predictive capabilities, PHS provides unique and scalable approaches to detect and diagnose diseases while lowering overall healthcare expenses. PHS has been effectively used to mimic a wide range of health outcomes, including contagious diseases, the flu, and asthma care. Machine learning has just lately begun to be employed end-to-end in PHS health. Machine Learning technology advancements offer new opportunities for next-generation data-driven medical decision-making. ML algorithms can be developed to collectively and automatically digest massive online available data about health events to make data-driven predictions for pandemics. A Machine Learning system can automatically evaluate numerous data sources; reveal valuable hidden ideas from limited records to assist health professionals in understanding the data easily and succinctly, and create healthcare policies and decisions. Studies involving machine learning in PHS usually fall into two categories: predicting pandemic time-series cases or classifying the collected dataset into different categories such as disease or non-disease.

We point the reader to this review covering popular machine learning algorithms used for social media-based public health surveillance.

2.2. Review Progression

A systematic approach is used to identify the various machine learning approaches used by the social media-based surveillance systems. Firstly, we extracted 1240 research articles that published studies related to our research from various scientific digital libraries and inspected them. The scope of this paper involves studies of the social media-based surveillance systems that predict the disease in real-time or near real-time using machine learning approaches. The selection criteria for the research articles were set to incorporate papers published in the year 2010–2018, based on the few keywords such as surveillance, outbreak, health, disease, and social media in the abstract and title of the paper. These keywords serve as the foundation for the subsequent research strategy. Furthermore, the combination or permutation of the keywords mentioned above is employed to investigate the research. Some boolean procedures are also

employed for a more thorough examination. The following scientific databases were explored to provide a comprehensive bibliography of research papers on social media-based surveillance systems in the healthcare domain:

• ACM Portal • IEEE Xplore • Science Direct • PubMed

It is challenging and crucial to select articles/findings that are extremely relevant to the research work. If the abstract or title or both explain social media or web-based surveillance, then we considered them for further research; otherwise, they were rejected. The next step we performed was to consider the articles that had utilized machine learning approaches in their methodology.

Extracting relevant information from a certain subject is a difficult and complex task. It includes mining techniques, approaches, theories, datasets, results, limits, publication year, author credentials, journal/conference reputation, and the future scope of the specific paper.

In a nutshell, by using the aforementioned keywords on the 4 scientific databases, a total of 1024 articles are identified. After examining whether the abstract/title of the article is related to social media or web-based surveillance in the healthcare domain, only 148 papers were chosen. Furthermore, only 26 out of 148 articles employed machine learning techniques on online collected data. As a result, in the literature review, all of these studies are thoroughly analyzed, along with their positive and negative characteristics. Figure 2.1 describes the steps followed in selecting the relevant articles for this study.

2.3. Literature Review: Machine Learning Methods used by Social Media based Health Surveillance Systems

This section explains the most commonly used machine learning-based classification methods to analyze health-related text from social media platforms. In recent years, machine learning has received much attention, particularly for recognizing patterns in pictures or raw data. M. Bates [157] discussed how advances in machine learning enable epidemiologists to mine a large amount of digital data. M. Conway, D. O'Connor [28] discussed the use of natural language processing and machine learning in combination with social media platforms to aid in the study of large datasets for population-level mental health research. Among the several methodological variants of machine learning, certain architecture is particularly popular. For example, we discovered that the precision of the k-Nearest Neighbor (k-NN) classifier was superior to that of various other machine learning classifiers, including NBM Modal, NB, and SVM [49], when it came to categorizing tweets into two classes—real instances of allergy or awareness tweets. Similarly, K. Lee, A. Agrawal, A. Choudhary, et al. [44] showed that when compared to other classifiers such as NB, RF, and SVM, the highest text classification performance was achieved using Multinomial Naive Bayes Modal with F-measure of 0.811. The author [43] demonstrated a model that used a multilayer perceptron with a backpropagation algorithm on Twitter data to predict the weekly

status of the ILI-infected population in the United States. Several supervised machine learning algorithms were investigated to detect personal health experience tweets [107]. The deep gramulator approach was used to improve precision when applied to independent test sets.

The unsupervised classification algorithms do not require labeled data sets to predict the output like supervised algorithms. Because of this reason, the unsupervised classification methods seem to be a more attractive alternative in the process of analyzing the text, but they could be more challenging in achieving a similar accuracy as supervised methods. The same can be observed when L. Sousa, R. de Mello, D. Cedrim, et al. [50] performed the classification of tweets using supervised and unsupervised methods. They concluded that topic modeling (LDA), one of the unsupervised techniques, presents less control over the content of topics in comparison to a traditional classifier, particularly on a naturally noisy media channel. Hence, Multinomial Naïve Bayes, a supervised classification approach, was considered for classifying the Twitter content. The following sections represent the commonly used Machine Learning methods for health-related text classification in selected papers.

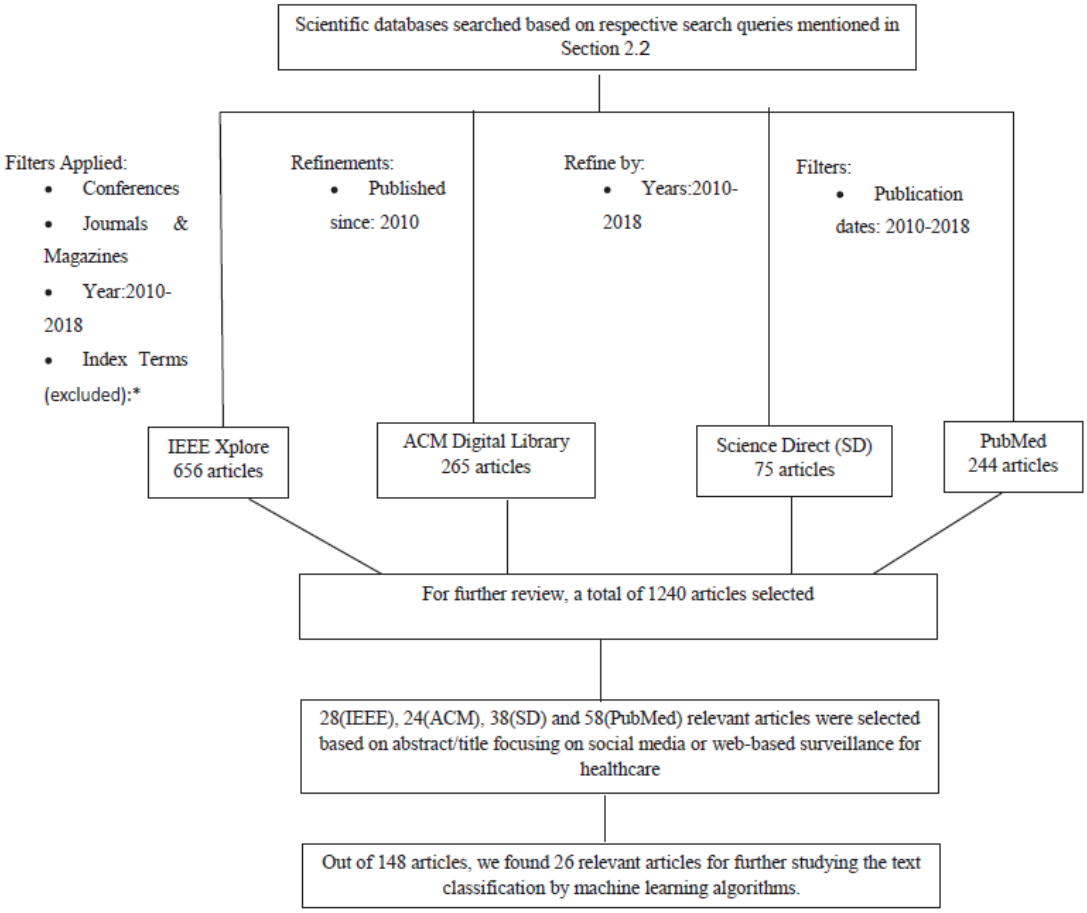


Figure 2.1 Search Methodology for the Selection of Relevant Articles

2.3.1. Support vector machine (SVM)

SVM is a popular binary classifier built upon the concept of hyperplanes to separate observed features into two classes. This approach transforms the original input data into a higher dimensional space using a nonlinear mapping. A linear optimal separating hyperplane is searched within this new dimension to minimize the distance between hyperplane points and maximize the margin between the classes [158]. It has been known for its superior performance in text classification with word features. Although the performance of classification algorithms highly depends on the input parameters and application, yet for binary classification tasks, SVM was observed to be highly suitable. V. K. Jain and S. Kumar [159] reported the SVM classifier as a good performer in terms of accuracy in predicting the class of tweets (disease-related tweets/irrelevant tweets). Similarly, the SVM algorithm was able to achieve an accuracy of 90.09% when tweets were classified as infodemiological or non-infodemiological [51]. N. Yang, X. Cui, C. Hu C et al. [70] were able to classify 'sick microblog' and 'not sick microblog' posts using the SVM classification model. They also showed that time consumption by SVM for the classification task did not get much affected when the micro-blogs needed to be arranged increased by 100 times, though there is a huge increase in the time consumption by k-NN to complete the classification task. SVM turned out to be the best tweet classification method when compared with other machine learning techniques and has been utilized to classify social media data on a range of physical health issues: [49,107,160–164].

2.3.2. Naive Bayes (NB)

It is a classification algorithm for binary and multiclass classification problems. This algorithm makes a naive assumption that there are no predictors, i.e., features are independent of each other, and one feature's impact on predicting class does not depend on the presence of another feature [165]. This method is based on the Bayes Theorem, shown below:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{2.1}$$

H: Hypothesis that data X belongs to a specific class C

X: Data with the class yet known

$P(H|X)$: Posterior probability of hypothesis H conditioned on X

$P(H)$: Prior probability of hypothesis H

$P(X|H)$: Posterior probability of X conditioned based hypothesis H

$P(X)$: Prior Probability of X.

V. K. Jain and S. Kumar [159] classified datasets into mosquito-borne disease-relevant and irrelevant tweets using SVM and Naive Bayes, and further relevant tweets were classified into three classes: symptoms, fear, and prevention using the same classifiers. V. Kumar and S. Kumar [160] performed the classification task on tweets to differentiate swine flu-related text from noise or irrelevant tweets using various ML techniques such as Decision Tree, Naive Bayes, SVM, and Random Forest. They considered every relevant swine flu-related word as a feature and found both Naive Bayes and SVM provided the best classification result, with an F-measure of 0.77. The Naive Bayes classifier performed best when dengue-suspected tweets were classified as irrelevant or relevant, considering emojis, location information, unigrams, bigrams, and trigrams [114]. Naive Bayes is a popular classification algorithm and is used by many authors for text classification: [46,49,50,93].

2.3.3. Multinomial Naive Bayes (NBM)

Multinomial Naive Bayes networks, a variant of Naive Bayes networks, are better for text documents. Multinomial networks consider the frequency of words, and underlying probability calculations are adjusted accordingly, while NB networks do not consider the frequency count.

X. Ji, S.A. Chun, and J. Geller [163] used different ML methods (NB, NBM, and SVM) to classify tweets as either personal or news-related. They further classified personal tweets into two categories: negative or neutral tweets. Among all the methods used for classification and sentiment analysis, NBM accomplished overall the best outcome and turned out to be better than the other two machine learning classifiers regarding the time consumed to build the classifier. Similarly, K. Lee, A. Agrawal, and A. Choudhary [44] developed a real-time allergy surveillance system that distinguished tweets as either positive or negative. A positive tweet mentions having allergy symptoms to an individual. And the negative tweet represents advertisement, news, or general awareness of allergies. They showed that the best text classification performance was obtained using NBM with the highest F-measure when compared to the other classifiers such as NB, RF, and SVM. Similarly, this classifier gave the best recall and f-measure of 0.859 and 0.857 when the tweets were classified among the multiple allergy hidden classes [49].

2.3.4. k-Nearest neighbor (k-NN)

K-Nearest Neighbor is an instance-based statistical analysis method to perform classification. Its implementation requires an integer k, training data set, and a measure closeness metric. The given training set is used as an input vector to form different regions for different classes.

When given an unlabeled object, this classifier searches for k training sets in the pattern space nearest to the unlabeled object. These training sets are the k “nearest neighbors” of the unlabeled object. “Closeness” of the sets is defined with a distance metric, such as Euclidean distance, which is given by:

$$dist\ (X1, X2) = \sqrt{\sum_{i=1}^{n}(x1i - x2i)^2} \quad (2.2)$$

Where $X1 = (x11, x12, \ldots, x1n)$ and $X2 = (x21, x22, \ldots, x2n)$ representing two objects or points [165,166]. Nargund K, Natarajan S. [49] used k-NN alongside Naive Bayes, SVM, and Naive Bayes Multinomial to identify and recognize messages reporting and discussing different types of allergies. They noted that k-NN has better precision than other approaches in identifying and assigning tweets as either actual incidents of allergy or awareness tweets. This classifier was observed as having a good precision for text classification than other classifiers, such as k-means, but not the best [70]. Other research papers that have used the k-NN are: [107,167].

2.3.5. Logistic Regression (LR)

LR is a statistical technique for analyzing a dataset for a binary classification problem [168]. It helps discover the relationship between a dependent binary variable and at least one independent variable. Each independent variable is multiplied with weights and summed up. This outcome will sum up to a sigmoid function to get the result in the range of 0 and 1. The values below 0.5 are considered 0, and those above 0.5 are considered 1. In this manner, optimization techniques aim to find the best regression coefficients and weights. Logistic regression is mathematically constrained to produce probabilities in the range [0, 1]. Also, it can converge on parameter estimates relatively easily.

Along with different classification algorithms, logistic regression is also preferred for the data classification task. For instance, the logistic regression gave a better recall and F1 measure than SVM in classifying asthma relevant and irrelevant tweets [161]. Logistic regression showed excellent precision for analyzing personal and non-personal tweets [107]. The research papers we have reviewed also utilized logistic regression classifiers [153]. Maximum Entropy classifier, sometimes called Multinomial Logistic Regression [169,170], is also used for the text classification task. Tweets related to illness were identified using Maximum Entropy [35]. Another study that used Maximum Entropy for tweet classification includes [114].

2.3.6. Decision Tree (DT)

A Decision Tree is a flowchart-like tree structure, where each internal node represents a test on a feature, each branch denotes the result of a test, and each terminal node holds a class mark. Attributes values of an unlabeled sample, X, are tested against the decision tree to predict its class. A unique path is traced from the root (topmost node) to a terminal node based on attributes' values, which holds the predicted class for the unlabeled sample [165,171]. DTs are easy to assimilate and have good accuracy. They can handle real-valued items, categorical features items, and items with a mixture of both. They are flexible enough to handle items with some missing features. Unfortunately, decision trees are poor at handling changes as

a minor change in input data may lead to massive changes in the constructed tree. They are good at naturally supporting classification problems with more than two classes and capable of handling regression problems. Finally, once constructed, new items can be classified quickly.

J-48-Decision Trees classifier performed well in predicting positive and negative tweets related to personal health experiences [167]—similarly, R. A. Calix and A. General achieved an average result using Decision Trees classifiers for classifying Personal health experience tweets [107]. Even the DT classifier was experimented for distinguishing tweets related to swine flu [160].

2.3.7. Deep Neural Network

A standard neural network (NN) consists of many simple, connected neurons, whereas a Deep Neural Network (DNN) employs a deep architecture in NNs with a certain level of complexity and an increased number of layers in a single layer [48, 49]. In the past few years, Deep Neural Networks (DNNs) have gained much popularity in text classification. DNN classifiers outperform every other conventional classifier experimented with, such as IB1-k-Nearest Neighbor, J48-Decision Tree, LR, and SVM, when tweets were classified as personal and non–personal health experience tweets [33]. CNNs have achieved remarkable performance in computer vision and deep learning. It is a class of Neural Networks that is proven very useful in text processing, image recognition, and classification. Recently, CNNs have been actively exploited for text classification in the health domain. J. Du, L. Tang, Y. Xiang, et al. have also used different types of DNN, i.e., CNN and Bidirectional Long Short-Term Memory (Bi-LSTM), along with other ML approaches for measles-related tweet classification tasks, where CNN has shown a remarkable performance [98]. To study the prediction of chickenpox and eliminate delays in disease reporting of existing surveillance systems, S. Chae, S. Kwon, and D. Lee [172] optimized parameters of DNN and LSTM (a special kind of RNN) algorithms.

2.3.8. Random Forest (RF)

Random Forest is an ensemble learner which improves the accuracy of the model by combining a collection of decision tree classifiers (forest) to generate the aggregated result. It uses Classification and Regression Trees (CART) methodology to grow the trees. At each node, attributes are randomly selected to determine the split and generate individual decision trees. The values of the random vector sampled are responsible for determining each tree. During classification, votes are cast by each tree, and the class with the maximum votes is returned [32].

RF can handle multidimensional data and is capable of estimating missing data. It also considers the importance of the variables used in classification. RFs consider many fewer attributes for each split, and

they are efficient on vast databases. The Random Forest approach is used along with other conventional machine learning approaches for social media text classification, such as [44,50,107].

Some authors have explored other popular machine learning approaches for text mining, such as k-means [70], clustering [46], etc. Dai X, Bikdash M, and Meyer B. [46] proposed a word embedding-based clustering technique to classify health-related tweets. A tweet can be grouped on the grounds of similar words and can be classified based on the similarity measure. They compared their proposed method with the Naive Bayes classifier and found the former is superior to the Naïve Bayes method. Topic modeling, a well-known semantic clustering algorithm, has indicated a valuable outcome for classification [50]. Table 2.1 represents the machine learning techniques adopted for the classification of health-related text and the social media sources used, along with corresponding research papers.

Table 2.1 Summary of Machine Learning Classification Approaches used in Health Surveillance Systems based on Social Media

Article Citation	Year	Health Topic	Machine Learning Approach	Social Media Data Source
[172]	2018	Chickenpox	DNN, RNN	Internet Search Query
[98]	2018	Measles	CNN, RNN, SVM, k-NN, NB, RF	Twitter
[50]	2018	Mosquito-borne diseases (Dengue, Zika, Chikungunya)	SVM, NB	Twitter
[173]	2018	Dengue	SVM, k-means	Twitter
[107]	2017	Personal-Health Experience	SVM, LR, k-NN, DT, DNN	Twitter
[41]	2017	MERS, Ebola	SVM,NB.LR	Twitter
[167]	2017	Personal-Health Experience	k-NN, DT, LR, SVM, DNN	Twitter
[159]	2017	Mosquito-borne diseases (Dengue, Malaria, Chikungunya)	SVM, NB	Twitter
[46]	2017	Flu	NB	Twitter
[51]	2017	Dengue, Typhoid	SVM	Twitter
[35]	2016	Influenza	ME	Twitter
[114]	2016	Dengue	SVM, NB, ME	Twitter
[164]	2016	Flu	SVM	Twitter
[161]	2016	Asthma	SVM, LR	Twitter
[49]	2016	Allergy	SVM, NB, NBM, k-NN	Twitter
[162]	2016	Influenza	SVM	Twitter

[96]	2016	Health Related	SVM, NB	Twitter
[170]	2016	Flu/Influenza, H1N1	LR	Twitter
[160]	2015	Influenza-A (H1N1)	NB, SVM, RF, DT	Twitter
[44]	2015	Allergy	NBM, NB, RF, SVM	Twitter
[78]	2015	Asthma	ANN	Twitter
[174]	2014	ILI	SVM	Twitter
[70]	2014	Flu	k-means, k-NN, SVM	Sina Microblog
[55]	2013	Influenza	SVM	Twitter
[163]	2013	Personal health	NB, NBM, SVM	Twitter
[175]	2013	Health Related	SVM,NB,NBM,RF	Twitter

Among all the types of social media platforms studied by the selected articles, SVM was the most used classification technique among the recently used Machine Learning based classifiers. In addition, it was observed that SVM was the most promising classifier when the data needed to be classified between classes.

This chapter aims to study the trends in surveillance systems using social media and machine learning algorithms in public health. We have noted that social media-based surveillance systems show superiority compared to traditional surveillance systems. Also, we observed some research gaps and limitations while reviewing the selected papers that, when addressed and resolved, will lead to the development of new technologies and new capabilities for public health research.

2.3.9. Existing State-of-the-art methods

There has been considerable research on using ML to improve public health surveillance systems. There has definitely been an explosion of public health data, as well as an explosion of interest in applying AI. Because ML is better suited to find structures within complex data sets than traditional statistical methods, their application can surely improve public health surveillance systems. In past years, DL models such as CNN [176–178], LSTM [178,179], BERT [179,180], etc., are widely used in learning feature representations and in many tasks, can provide better performance than the ML models. This section discusses some recently used state-of-the-art DL techniques to examine health-related data from social media platforms.

2.3.9.1. Extreme Gradient Boosting (XGBoost)

XGBoost is a type of Gradient Boosting. Gradient Boosting allows the use of any differentiable loss function. It is a reliable and efficient off-the-shelf approach that can be applied to both regression

and classification problems. XGboost is a decision tree-based model and has become one of the popular DL algorithms in the past few years. XGboost is robust against overfitting as they use L1 and L2 regularization methods, which are not present in Adaboost and GradientBoosting techniques. In a research, M. Kamal, S.U. Rehman Khan, S. Hussain, et al. proposed a novel machine learning algorithm to identify the patient's mental disorder based on Reddit posts [181]. The proposed approach based on XGBoost for accurate data categorization into four different categories of mental disorders has outperformed other ML techniques such as NB and SVM. Another recent study has employed the XGBoost classifier to detect health-related public health posts [182]; however, it has not achieved the highest accuracy with every feature representation method. Similarly, J. Kim, J. Lee, E. Park, et. al employed XGBoost and a CNN to classify depression- and anxiety-related textual data, achieving an accuracy of 77.81% on anxiety-related data and 75.13 % on depression-related data [183].

2.3.9.2. Recurrent Neural Network (RNN)

RNNs are the most significant architecture of this period. RNNs are a common choice for any kind of sequential data; these structures are designed to extract information about sentence structure, including latent correlations between context words. An RNN model's input is often a phrase represented by a sequence of word embeddings, with each word entering the model one at a time. LSTMs are the most often used RNN variation because they address the vanishing gradient issue that traditional RNN architecture experiences [184]. LSTM is a kind of neural network with a "memory unit" capable of retaining knowledge in memory over long periods of time, allowing it to learn longer-term dependencies. The LSTM model was integrated with CNN to categorize suicidal and non-suicidal words in social media forums. The proposed approach was tested against LSTM, CNN separated, and other ML algorithms such as RF, NB, XGBoost, and SVM. The authors discovered that the proposed model enhanced accuracy with 93.8%, recall with 94.1 %, precision with 93.2 %, and F1 score with 93.4 % [178]. Another study used LongTerm Short Memory to identify depression tweets in Malaysian urban centers at the start of COVID-19 to help individuals, carers, and even medical experts recognize language indicators that point to symptoms of mental health deterioration. The implemented model achieved an accuracy rate of 94%, with precision, recall, and F1 score of 0.94, 0.96, and 0.95, respectively[185]. R. Biddle, A. Joshi, S. Liu, et al. used sentiment analysis to identify figurative health mentions [186]. Word embeddings are used to encode the words in a tweet, which are then put into a Bi-LSTM and a sentiment detection module, respectively. The results of the Bi-LSTM and sentiment modules are combined and fed into a softmax classifier, determining if the input data include figurative health references. The authors discovered, however, that the model fails to capture the complete contextual information of the disease term, and overall performance does not increase.

2.3.9.3. Convolutional Neural Network (CNN)

Convolutional Neural Networks are multilayered, feed-forward neural networks that use perceptions for supervised learning and information dissection. It is utilized primarily with visual information. However, neural networks are not limited to image recognition. Those were formerly utilized in text data analytics.

J. Du, Y. Zhang, J. Luo, et al. [187] used deep learning-based approaches to evaluate multiple techniques for identifying suicide-related mental stresses from Twitter. The results suggested that these strategies outperform ML algorithms. They improved the accuracy of classifying suicide-related tweets with an F1 score of 83% and a precision of 78% using CNN, outperforming Extra Trees, SVM, and other ML methods. In another paper, Sawhney, R.R. Shah, V. Bhatia, et al. [188] examined feature selection using the Firefly algorithm to develop an effective and robust supervised technique for detecting suicide risk in tweets. After applying several ML approaches, CNN-LSTM produced the greatest results in terms of accuracy, precision, recall, and F1 scores in particular datasets. S. Wang, J. Du, L. Tang, et al. proposed a multi-task CNN model to classify measles-related tweets in different categories based on their characteristics [189]. Six comparable traditional ML models were examined as baseline models, including Extra Trees, SVM, RF, LR, AdaBoost, and Gradient Boosting, and deep learning models outperformed conventional machine learning methods. Another dual CNN approach proposed by L. Luo, Y. Wang, and H. Liu research used the COVID-19 public health mentions dataset with over 11,000 annotated tweets to solve the issue of class imbalance [177]. In the dual CNN structure, an auxiliary CNN offers supplementary information to the primary CNN to identify public health tweets more efficiently. The experiment demonstrated that the presented framework might mitigate the effects of class imbalance and produce favorable outcomes.

2.3.9.4. Transformer-based Models

The transformer [190], in contrast to RNN-based models, is an attention-only deep neural network that learns global-level information from all inputs and selects the most significant information. Recently, transformer-based deep learning techniques such as Bidirectional Encoder Representations from Transformers (BERT), Robustly optimized BERT approach (RoBERTa), etc., have shown significant performance in various NLP tasks. These models significantly improve NLP tasks by estimating missing words in the text. Because they evaluate every phrase without bias, they perform better at comprehending the context of words than earlier ML approaches such as word embedding algorithms.

In recent years, BERT has been the most significant development in the domain of NLP, yielding cutting-edge outcomes in various NLP tasks. BERT has a multi-layer bidirectional Transformer encoder architecture [191]. It employs a Transformer, an attention mechanism that learns contextual relations between words in a document. It is designed to pre-train deep bidirectional representations from

unlabeled text sources by conditioning across both layers in both the left and right contexts. BERT is also trained on a secondary task, Next Sentence Prediction, in which it is given two sentences and must determine whether the second sentence follows the first in a binary form to allow the model to better learn sentence relationships.

Y. Liu, M. Ott, N. Goyal, et al. discovered that BERT was severely undertrained and proposed an improved way of training BERT models called RoBERTa, which can match or even outperform the performance of existing BERT approaches [192]. RoBERTa , too, was developed using a roberta-based modification of the huggingface [193] library's pre-trained model, whose architecture comprises BERT-base architecture along with 12 layers, 768 hidden, 12 heads, and 125M parameters.

DistilBERT, a distilled or approximate version of BERT, is a technique to pre-train a smaller general-purpose language representation model, which can then be fine-tuned with high performance on a variety of applications [194]. It is a lightweight deep learning model that has about half the total number of parameters as BERT and retains 95% of BERT's performance.

Like BERT and RoBERTa, XLNet [195], [195], a generalized autoregressive language model, is another pre-trained transformer-based transfer learning model. XLNet returns the joint probability of a series of tokens based on the transformer architecture with recurrence. It has state-of-the-art results in almost all the language modeling benchmarks. XLNet adopts a permutation language model to overcome the discrepancy since BERT and its derivatives ignore the dependency between the masked locations and undergo a pre-train fine-tune gap.

K. Zeberga, M. Attique, B. Shah, et al. designed a robust framework BERT-Bi-LSTM for detecting mental health concerns. Using DL techniques such as Bi-LSTM, BERT, and knowledge distillation based on user-generated social media information. The proposed framework improves the accuracy of smart healthcare systems in detecting mental-health disorders, including depression and anxiety [179]. During COVID-19, A. Murarka and I.B.M. Raleigh collected data on mental health using the Reddit API. They applied the RoBERTa model to analyze mental illness on social media [196]. G.S. Bajaj, H. Yadav, H.S. Sahdev, et al. employed transformer-based classification models such as BERT, RoBERTa, XLNet, Longformer, and DistilBERT on the COVID-19 data collected from Reddit and found RoBERTa produced the best results when trained on the collected dataset [191]. In recent research, the authors used posts from the social network Twitter to identify persons who have or have had an eating issue [180]. They compared six BERT-based prediction models. The top scoring model among the trained models was the one known as RoBERTa with 87.5% accuracy, followed by DistilBERT with an accuracy of 86.3%. X. Wang, S. Chen, T. Li, et al. investigated the potential of state-of-the-art DL approaches for depression risk prediction using microblogs [197]. DL methods with pre-trained language representation models, such as BERT, RoBERTa, and XLNet, were investigated and compared on a manually annotated

benchmark dataset for depression risk prediction. BERT outperformed the other two deep-learning algorithms, with a microaveraged F1 score of 0.856. RoBERTa scored the best performance on depression risk at levels 1, 2, and 3, with a macroaveraged F1 score of 0.424, setting a new benchmark result on the dataset.

2.4. Research Gaps and Limitations

After performing the profound, systematic literature on the selected studies, most concealed facts have been identified. Research gaps have been discovered as follows:

Limited inclusion of non-medical determinants of health: The non-medical determinants such as socioeconomic position, health system metrics, environmental pollution, climatic conditions, demographic variables, etc., have an important influence on the dissemination of contagious diseases and pandemics. It has been observed that most of the studies have worked only on historical cases for pandemic prediction. The influence of non-medical health determinants was not considered while forecasting the trends of infectious diseases. This study has included the influence of COVID-19 pandemic-related multisource data in pandemic prediction. The outcome is also compared with the traditional methods, and the results show that the inclusion of nonmedical health determinant factors has improved the prediction performance.

Limited incorporation of related input features: Social media content consists of users' sentiments, opinions, reviews, etc. It is in the content of the online data that a business will end up understanding more about their audience, specifically on sites like Instagram, Facebook, and Twitter. Many applications, including marketing, time-series prediction, etc., make extensive use of social data. Yet, we found that there is a scope for incorporating the input features such as sentiment content, comments, locations, etc., of users' posts along with the text content for the better analysis and prediction of health events and other related events.

Different weightage to different categories: In the literature review, the authors have categorized the online collected data into various classes/categories. Though different content has different influences on the spread of the event yet, the papers have assigned the same weightage to all the categories contributing to an event. Therefore, the forecasting accuracy of an event can be improved by classifying the users' posts into different categories, such as health, news, ads, etc., and assigning them different weights.

Inclusion of video data: The amount of information being uploaded onto social platforms, such as YouTube, Instagram, Facebook, etc., can be a good source for finding informative and trustworthy online information. Yet, there is limited use of the video content generated on social media platforms for public health surveillance. Therefore**, t**he text and image analysis approaches can be extended to video content analysis to enhance disease prediction performance.

Limited use of unsupervised Learning: Most prediction models have implemented supervised machine learning methods for text classification in the retrieved papers. Although the contribution of supervised approaches to progress in text mining tasks is pretty well-known, this is less true for other types of learning, especially self-supervised and unsupervised learning. We noticed that very few selected research papers had explored unsupervised ML methods or deep learning approaches for text classification. Supervised learning, specifically classification techniques, was mostly used to construct the ML model for PHS; more studies should be conducted utilizing other methods such as clustering and/or association in unsupervised learning techniques. Also, we observed there is a wide scope of improvement in the area of topic modeling to get more accurate results.

Lack of feature extraction and selection techniques: The literature lacks the use of feature extraction and feature selection techniques. The lack of feature extraction and selection is one of the reasons for the insufficient accuracy of the time-series prediction models.

Limited number of social media platforms explored: However, discussions and messages of health events or crises occur on multiple internet-based platforms, yet most papers have relied on a single web data source for surveillance purposes. Therefore, there is a scope for implementing predictive models that can simultaneously use various social media and online platforms to get accurate and timely predictions of epidemics.

Estimating the severity of the outbreak: We noticed that very few studies have worked on the magnitude and severity estimation of the outbreak while performing health surveillance. Once an outbreak has begun, knowledge about the potential severity in real-time can help public health authorities to respond effectively. Nonetheless, predictive modeling of the magnitude of a pandemic outbreak is more significant than retrospective investigations.

2.5. Chapter Summary

This chapter presents systematic literature that profoundly discusses the machine learning algorithms studied for public health surveillance. A systematic study plan has been described that selects only the relevant studies related to social media-based public health surveillance using Machine Learning techniques. In the entire study, 1024 papers were identified for study, but eventually, a total of 26 articles were found relevant for studying the ML-based text classification algorithms on social media-based health data. Thus, this chapter presents relevant and up-to-date literature about public health surveillance using machine learning techniques.

Chapter 3 AN EPIDEMIC MODEL FOR TIME SERIES PREDICTION USING MACHINE LEARNING

Mathematical models may be a valuable tool for public health officials in epidemic control, potentially leading to a large decrease in the number of infected individuals and fatalities. Furthermore, mathematical models may be used to generate long- and short-term forecasts, allowing decision-makers to maximize potential control techniques, including vaccination programs, lockdowns, and containment measures. Several models for explaining the temporal evolution of epidemics have been proposed, such as Susceptible-Infected-Recovered (SIR) models [198], sub-epidemics wave models [199], Richards models [200], logistic models [198], generalized growth models [201]. Traditional epidemic models assume that all individual coronaviruses infect at the same rate. These models' forecasted outcomes can only provide general trends, limiting their usefulness. The government's preventive and control initiatives have substantially influenced the virus's eradication. Transparent reporting of the pandemic, increasing preventative knowledge among individuals and implementing measures to prevent and control infection have all aided in the virus's elimination. Clearly, epidemiological data alone are insufficient to properly predict disease transmission. As a result, a data-driven pandemic model is essential for public health emergencies. Given the recent outbreak of COVID-19, this chapter proposed, Deep-SIQRV (Susceptible–Infected–Quarantined–Recovered–Vaccinated), an epidemiological model that predicts the infection rates for studying development trends and transmission laws. Second, the model takes into account the impacts of preventive and control efforts, as well as the growth in public prevention awareness, by including data from online news and social media posts. The proposed model is a novel combination of traditional mathematical disease modeling with the equation-based SIQRV model, LSTM embedded with the PAN-LDA module, and Evolutionary Strategies (ES). Section 3.1 explains the components and framework of the proposed model. Section 3.2 talks about the methodology. The next section, i.e., section 3.3, presents the experimental results. Section 3.4 wraps the summary of the proposed work.

3.1. Deep-SIQRV Model

In this chapter, we investigate the Deep-SIQRV model, which accounts for the fact that vaccine protection diminishes over time and that vaccinated persons may become ill, and then we expand the model's structure by including alternate infection rates.

3.1.1. Components of the Model

The proposed model consists mostly of four components, which are listed below:

SIQRV Model: The SIQRV compartmental model is an infectious disease epidemic model in which the population is divided into five compartments: those susceptible to disease, S, those infected with the disease, I, those quarantined Q, those recovered from the disease R, and those vaccinated V. Throughout this research, we have considered that people who have recovered can become susceptible again. Infected persons include both symptomatic and asymptomatic infected people. Furthermore, the impact of vaccine protection decreases with time, and the vaccinated individual might revert to being susceptible.

Figure 3.1 shows the diagram of the SIQRV model network and the transformation relationship between the individual component. In this case, infected individuals include symptomatic and asymptomatic infected people. If a susceptible person is exposed to an infected individual, it gets infected with probability β. The susceptible individual is vaccinated with probability v, and the probability of death of a quarantined individual is μ.

Contrary to the conventional SI model, vaccination does not provide lifetime immunity to infection, and hence, vaccinated persons become susceptible again with rate s. Setting a city lockdown switch based on the death rate, infection chance β varies depending on the city condition (lockdown or not). Under strict government regulation, the lockdown infection probability β is significantly smaller.

The dynamic equations may be expressed as follows based on the propagation of the epidemic:

$$\frac{dS(t)}{dt} = -\beta S(t) - vV(t) + sV(t) + \gamma R(t) \tag{3.1}$$

$$\frac{dV(t)}{dt} = vS(t) - sV(t) \tag{3.2}$$

$$\frac{dI(t)}{dt} = \beta S(t) - \delta I(t) \tag{3.3}$$

$$\frac{dQ(t)}{dt} = \delta I(t) - rQ(t) - \mu Q(t) \tag{3.4}$$

$$\frac{dR(t)}{dt} = rQ(t) - \gamma R(t) \tag{3.5}$$

The probability of death may be calculated using the normalizing condition:

$$mortality\ rate = \frac{\gamma\mu\delta\beta}{\beta[(\delta+r)+\mu(1+\delta)]} \tag{3.6}$$

In this work, we focus on the mortality rate as a key element in determining the propagation of an epidemic. It not only displays the disease mortality rate during the coronavirus spread, but also the efficacy of disease prevention and control methods (e.g., increasing public awareness of infection spread will lower the infection risk, city lockdowns, and house quarantines [202–204]).

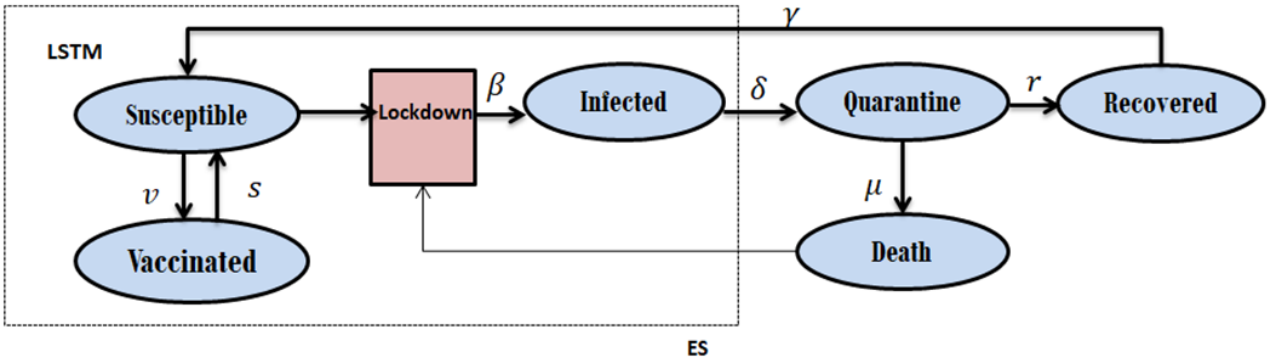


Figure 3.1 The Transformation Relationship Between the States of the SIQRV Model

Evolutionary Strategy (ES): Evolutionary Strategy is a nature-inspired optimization algorithm that employs a population of people that are assessed by a fitness function and influenced by mutation [205]. The ES employs a fitness function to assess a population of individuals distributed by mutation. It also has fewer hyperparameters than the Genetic Algorithm and is unaffected by settings with sparse rewards.

Long Short-Term Memory (LSTM): The LSTM is an improvement on the RNN [206] since it includes a processor that assesses whether or not the information is relevant, and the functioning component is referred to as a cell. In a cell, there are three gates: the input layer gate, the output layer gate, and the forget-gate. Both the input gate and the forget-gate affect the status of cells. Though, the input gate's purpose is to decide the relevant information to add up from the current step, whereas the forget gate role is responsible for selectively forgetting the information about cell states [207]. In order to emit information, the output gate works on the hidden layer. Figure 3.2 depicts the structural model of an LSTM unit. The goal of our work was to optimize the infection rate of COVID-19; because this is a task well-suited for the LSTM model, we employed it in our study.

Pandemic-LDA (PAN-LDA): PAN-LDA is a topic model that uses LDA to find features in a mixture of text, particularly news items and historical time data [208]. PAN-LDA features are used as supplementary input features for any machine learning algorithm to enhance pandemic time series prediction. We provide the posterior distributions utilized in collapsed Gibbs sampling for PAN-LDA and propose a methodology for using PAN-LDA in text and data mining to forecast pandemic time series.

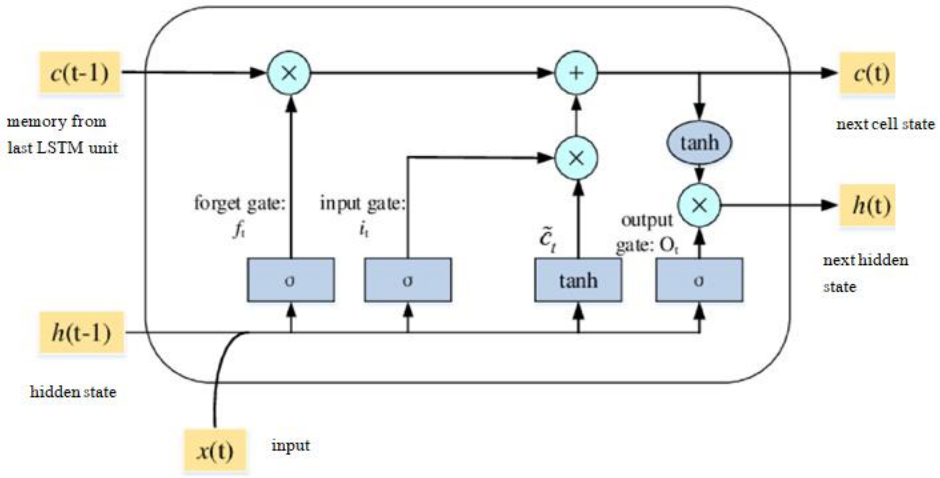


Figure 3.2 The Structure of an LSTM Unit

3.1.2. Framework of the Proposed Model

While assessing the development trend of coronavirus, most existing pandemic models consider the number of infected but non-quarantined individuals as the infection source for daily new infected cases in the future [209]. The infection rate of coronavirus-infected persons, on the other hand, differs during various times of infection [210]. Conventional models consider all infected people with COVID-19 to have the same transmission rate of infection and are inefficient in representing an epidemic's evolving trend [211]. This paper considers that under the prevention and control methods, most of the newly confirmed coronavirus cases have been infected by existing coronavirus-infected patients in recent days. On the other hand, asymptomatic infected people have the potential to spread the disease, which is challenging to identify. As a result, newly infected cases on day t and reported cases in the last k days to have a specific connection. Furthermore, the rate of infection in patients is directly linked to the timing of infection. The rate of infection in newly confirmed cases may change over time on day t during the last k days as a result of government efforts and media publicity that can slow the spread of the illness. Based on this assumption, this paper estimates the infection rate using the ratio of total infected cases at time t to total infected cases across various time periods prior to time t. The proposed model employed grouped multiparameter variables to evaluate the effect of confirmed coronavirus cases at various time intervals prior to time t on coronavirus infected cases at time t to predict the rate of infection of infected individuals at different times. This enhanced model is then utilized to investigate the development trend of the infectious disease.

Furthermore, the hybrid method of the Evolution Strategy and LSTM network is utilized to optimize the SIQRV epidemiological model to examine the infection rate deviation and estimate the number of infected cases. First, ES is utilized to calculate the infection rate transmission, with the fitness function

being the acceptable mortality rate which was the basis for city lockdown parameters. Furthermore, the bias for the transmission rate of infection rate in the pandemic model was assessed through the LSTM neural network, which is then coupled with the SIQRV model to predict the infected cases. In addition, we used the PAN-LDA model for extracting features from a large set of relevant news articles and social media posts data to assess the impact of government control measures, transparent reports by media, and increased public knowledge about epidemic prevention. The Deep-SIQRV model, which can forecast the infected cases based on development trends and transmission laws, uses the extracted features by PAN-LDA to update the infection rate deviation calculated by the LSTM network. The optimal model of disease transmission can be obtained by combining these techniques, which can be used to predict the number of infected cases. The framework of the Deep-SIQRV model is shown in Figure 3.3.

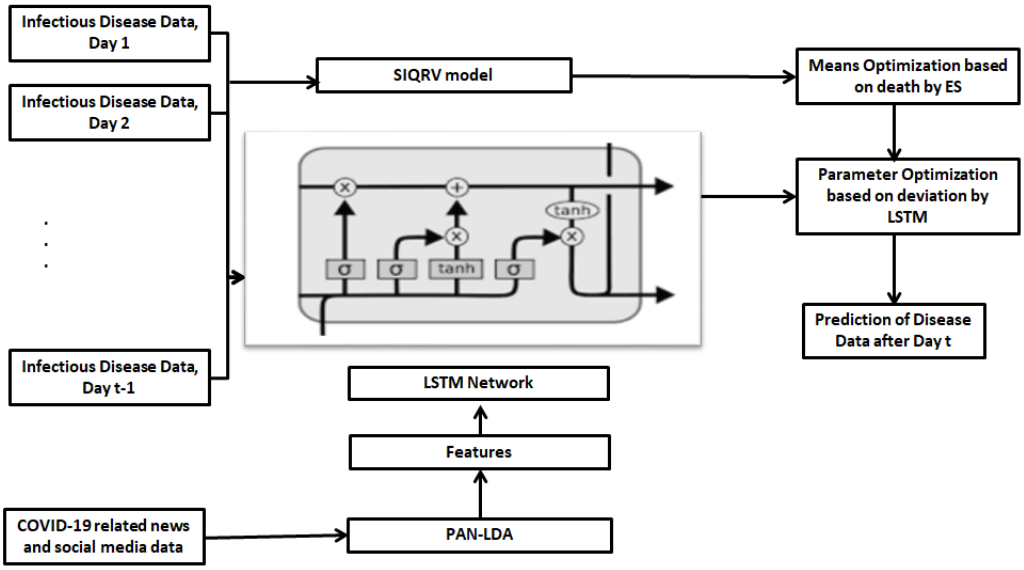


Figure 3.3 Framework of the Deep-SIQRV Model for COVID-19 Prediction

3.2. Methodology

3.2.1. Data Collection and Pre-Processing

Reddit Data

Social media is an excellent resource for gathering health-related data [212]. Reddit is a social media website that hosts discussions on various topics, including web content ratings. Users on this social networking platform may submit questions and comments about various topics, such as COVID-19, and reply to each other. It offers users a compilation of recent news, online content ratings, and user discussion on almost any topic. The posts are categorized by topics generated by internet users, known as

"subreddits," and include various topics. This research emphases on coronavirus–related comments from 8 different subreddits including r/indiacorona, r/CoronavirusIndia, r/covidIndia, r/COVID19, r/Coronavirus, r/COVID19support, r/nCoV, and r/CoronaVirus2019nCoV as a step toward creating this model. As a consequence, we evaluated 5,63,079 coronavirus–related Reddit comments. The data was gathered between March 1, 2021, and May 16, 2021.

Online News

We manually gathered English online news at regular intervals between March 1, 2021, and May 16, 2021. To extract the news articles, we used the Google search engine with the search keywords "covid", "coronavirus", "corona", "India", and "news". After reading the news items, content unrelated to COVID-19, duplicates, and inactive web links were deleted. After that, we examine the COVID-19-related information, sort it by date, and filter out case reports and relevant international news to identify the main elements of news information. One of the most important processes in preparing textual data is eliminating unnecessary words/data from the raw text, known as stop-words in Natural Language Processing. Furthermore, by eliminating stop-words, we reduce the dimensionality of the feature space. The bag-of-words representation of the documents on day t is fed into the PAN-LDA model, which is then transformed to feature vectors by the pre-trained PAN-LDA model. The text is encoded using the pre-trained last hidden layer of the LSTM. The 1024-dimensional news data and 1024-dimensional Reddit comments are combined to form a 2048-dimensional feature vector, each vector representing a part of the textual data. The recorded Reddit comments and news are categorized by date, and the feature vector is calculated by averaging the features of all textual data from that day.

Historical Data

Next, we collected historical data from the World Health Organization between March 1, 2021, and May 16, 2021 [213] and included three sorts of information: "Date," "Confirmed", and "Deceased". The vaccinated data is then obtained from [214].

3.2.2. Optimization of SIQRV Model Parameters

This paper used ES primarily to optimize model parameters and is divided into two parts. The first part explains how to use the ES to maximize lockdown conditions in a city. First, an initial value is established for the model, using permissible mortality as a criterion for assessment. According to regulatory standards, a three-day fatality rate of more than 0.045 is considered undesirable [215,216]. When the unacceptably high fatality rate is reached, the LSTM model parameter will be updated and saved as a basis for lockdown in a city.

Simultaneously, run the model with minimal mortality as the fitness function and use an evolutionary approach to adjust the neural network parameters. The optimal neural network parameters, or control measures, are generated using a twice evolutionary process. The LSTM method is also used to change other model parameters, and historical data are compared to model data, i.e., total confirmed cases and fatalities, to see if the model has attained minimal error. Further, the number of infected cases is predicted.

3.2.3. Infection Rate Calculation

Unlike the conventional SIR epidemic model, which uses an invariable infection rate to represent the infection probability, this model uses a variable infected rate. This study proposes a dynamic method for representing the spread of novel coronavirus disease.

In most conventional epidemic models, confirmed cases on a given day are considered to have been infected by the infected individuals over the past few days. These epidemic models use a specific number of analysis days and consider that the preceding k days influence the pandemic propagation. However, such models lack a detailed investigation of the epidemic transmission process. The majority of studies on COVID-19 revealed that the coronavirus observation time is 14 days [217]. This article considers that due to rigorous control and quarantine procedures, coronavirus-infected patients, after being quarantined, cannot infect the number of susceptible individuals. As a result, we may assume that the infected or confirmed patients in the previous 14 days are the sources of most of the daily new confirmed cases [209]. This study considered that a person infected on day t is most likely to be infected by an infected individual from day $t-1$ to $t-13$. In this article, the infection rate is examined at different intervals of time and analyzes whether the infection source of infected cases on day t are the cases from day $t-k$ to t-1. The equation (3.3) can be converted to:

$$\frac{dI(t)}{dt} = \beta_1(t,k)\sum_i^k \Delta I(t-i) - \delta I(t), \quad k = 1,2 \dots ,13 \tag{3.7}$$

The infection rate $\beta_1(t,k)$ indicates the association between the newly infected patients $I(t)$ on day t and the total infected patients $\sum_i^k \Delta I(t-i)$ in the last k days.

Furthermore, there is a strong correlation between the time of infection and the infection rate [218]. As a result, the transmission rate of infection for the newly infected individuals on a given day t may differ from those of the patients confirmed at other times in the last k days. In order to remove this disparity further, we assigned different weights α_i to daily new infected cases from day $t-k$ to $t-1$. The weighted total infected cases are then utilized to evaluate the infection rate, which is then used to model the pandemic. As shown in equation (3.8), the epidemic modeling is carried out:

$$\frac{dI(t)}{dt} = \beta_2(t,k)\sum_i^k \sigma_i(\Delta I(t-i)) - \delta I(t), \quad k = 1,2\ldots,13 \tag{3.8}$$

The above-proposed model studies the transmission relation between the infected cases over the last k days and the newly infected cases on day t. The transmission relationship is examined by accounting for the difference between the rate of infection of newly infected cases generated in the last k days versus infected cases on day t. First, a multiparameter epidemiological model is developed (3.8) by randomly initializing several groups of varying weights. The more accurate the model's prediction outcome, the more accurately the corresponding weights reflect the infection law. Lastly, by analyzing the weights allocated to different periods, the infection rate with a substantial impact on viral infection may be deduced. Based on (3.7) and (3.8), we can determine the association between the newly reported cases on day t and $t-13$ to $t-1$ through the value of σ_i .

We remove this disparity further by assigning different weights to the days with the most significant influence on newly infected cases on day t and the other days. Using equation (3.9), we carried out the multiparameter epidemic modeling. The number of examined days $\{t-i|\ i = 1,\ 2,\ldots,13\}$ were split into two groups, with set A recording the days with the most significant influence on newly infected cases on day t and set B recording the other days. As in, set A is given a weight of w_1, while set B is assigned a weight of w_2.

$$\frac{dI(t)}{dt} = \beta_3(t)\left(w_1\sum_{t_1\in A}\Delta I(t_1) + w_2\sum_{t_2\in B}\Delta I(t_2)\right) - \delta I(t) \tag{3.9}$$

where $\alpha_1|A|+\alpha_2|B| = 1$. The infection rate is calculated using (3.9).

3.2.4. PAN-LDA and Infection Rate-based LSTM Model for Trend Prediction

Deep neural networks can fit complex distributions, but they tend to overfit without enough supervision. Infection rate features are consistent over time since they are dependent on the growing percentage of each factor. On the other hand, epidemic models based on infection rates are unable to foresee policy changes or emergencies, nor can they be adjusted to account for short-term influences. As a result, we present the PAN-LDA features-based LSTM model to imitate existing policy and social media (Figure 3.4), ensuring both long-term and short-term stability. Our objective is to create a neural network architecture in which the upper layer is an LSTM for forecasting infection cases, and at each time point, the lower layer provides the LSTM with a latent topic distribution learned from the collected textual input. Two judgments were used to discuss the model parameters. When the model's death rate reaches an unacceptable level, the closed city's neural network parameters are utilized to process the parameters of the model. The mortality and infection rates are then compared to the real data. The optimal model parameters are eventually achieved by getting the minimal model error. In the proposed model, the real

infection rate is assumed to be $\beta(t)$, and the regressive infection rate be the exponential function, $\widehat{\beta}(t)$. We employ an LSTM to anticipate the deviation between the real and regressive infection rates. The deviation feature for prediction is the label of day t, which is set to $y(t) = \beta(t) - \hat{\beta}(t)$. As a result, the LSTM network can be used in conjunction with the SIQRV model. Figure 3.4 depicts the prediction flow diagram, where I_{Ri} is actual infected individuals on day i; $I_{\mathrm{P}i}$ denotes the infected cases predicted on i^{th} day and M_i represents the model on day i.

We integrate the PAN-LDA features mentioned in section 4.2 with the bias features to account for the impact of news and policies. Lastly, an LSTM model was used to analyze the contribution of coronavirus-related comments posted on Reddit and the news articles for public awareness and media publicity. We designed a neural network using two-layered LSTM and used the 50-dimensional GloVe vectors for pre-trained embeddings, trained on a large corpus of coronavirus–related textual data. However, by converting pre-trained embeddings, the processed text is converted to fixed dimension vectors. Furthermore, collected textual data can be represented as a sequence of characters with a corresponding dimension, resulting in a matrix [219]. We add a single-layer perceptron model with a fully connected layer. We employed the LeakyReLU activation function in the fully connected layer to convert the PAN-LDA and infection features into 64-dimensional vectors. This method assures that both features contribute equally to our model (Figure 3.4).

Given PAN-LDA features s_1 and infection features s_2, let W_1 and W_2 be the weights of the first two perception models. Let $\mathrm{g}(.)$ be the convolutional function and LeakyReLU, respectively:

$$f_1 = g(s_1; W_1) \tag{3.10}$$

$$f_2 = g(s_2; W_2) \tag{3.11}$$

The processed features f_1 and f_2 are concatenated to form the mixed feature f. At every timestamp t, let $f(t)$ be the mixed feature and h_{t-1} be the hidden state from timestamp $t-1$. The lstm function consists of the LSTM model and a fully connected layer that converts the hidden state to predict confirmed cases.

$$(x(t), h(t)) = lstm(f(t),\ h(t-1); W_l) \tag{3.12}$$

where W_l is the network's weight. And $x(t)$ and $h(t)$ are the output of the LSTM model and the new hidden state, respectively.

During training, we optimize using the Adam optimizer [220]. The mean-squared error is used as the loss function between prediction and label.

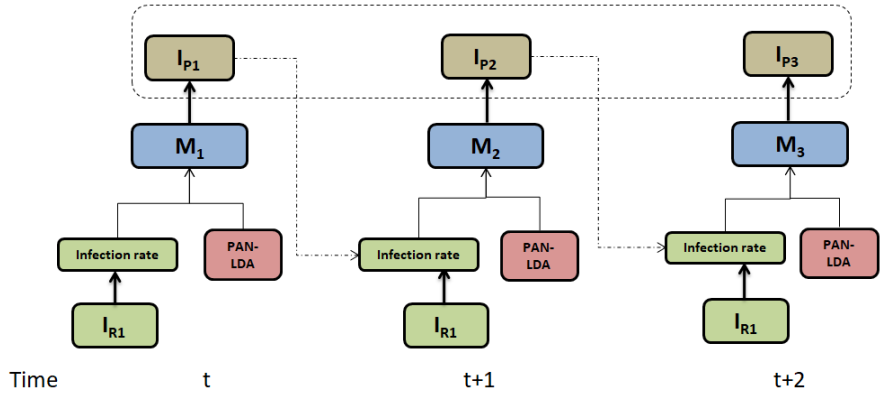


Figure 3.4 LSTM Neural Network

3.3. Experiments and Results

3.3.1. Infection Rate Calculation

Infection rates vary across individuals incubating at various infection time intervals [221]. Daily newly confirmed patients in the last k days may impact differently the infection rate of reported cases on day t. We explore the impact and temporal laws of epidemic spread using (3.8).

We investigate the association between newly confirmed COVID-19 cases in the preceding 13 days and newly confirmed cases on day t in Maharashtra, Kerala, Karnataka, and Delhi. When the weight distribution is taken into account, the curve of parameter t resembles a bell curve. As demonstrated in Figure 3.5, daily new cases from $t-9$ to $t-4$ days contribute more to new cases on day t, than from t$-$13 to $t-10$ and $t-3$ to $t-1$.

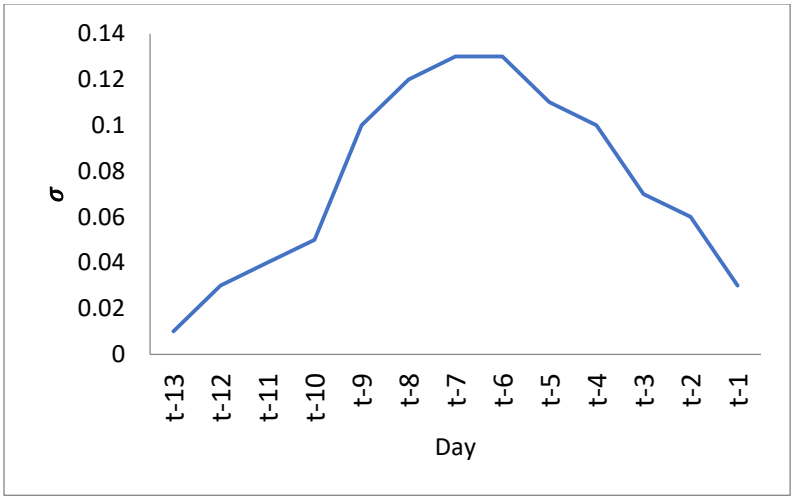


Figure 3.5 The Average Effect of Newly Infected Cases in Days (t – 13) to (t – 1) on Infected Cases on Day t in Maharashtra, Kerala, Karnataka, and Delhi

On fitting the estimated parameter $\beta_2(t)$ using equation (3.8), the distribution of σ_i reveals a pattern in which the value is modest on either side and high in the center. In the meantime, σ_i on day $t-13$ is near to 0, suggesting previously confirmed cases had a slight impact on cases confirmed on day t. Also, σ_i is larger on days $t-9$ to $t-4$, whereas σ_i is lesser on $t-13$ to $t-10$ days and $t-3$ to $t-1$ days. As a result, the average infection time is around 5.5 days.

We balance the estimated parameter $\beta_2(t)$ using a grouped multiparameter approach to avoid the underfitting and overfitting problems. According to (3.9), parameter w_1 is set as the weights of $(t-9)$ to $(t-4)$ days, and the parameter w_2 be the weight of $(t-13)$ to $(t-10)$ days and $(t-3)$ to $(t-1)$ days. Now, (3.9) can be transformed into:

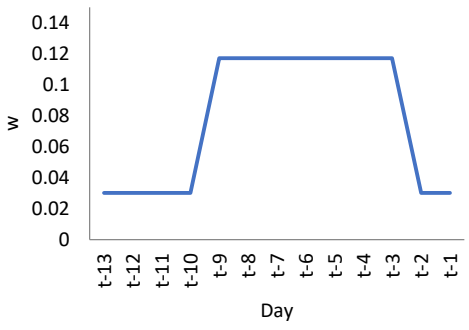
$$\frac{dI(t)}{dt} = \beta_3(t)w_1 \sum_{i=4}^{9} \Delta I(t-i) + \beta_3(t)w_2\left(\sum_{i=1}^{3} \Delta I(t-i) + \sum_{i=10}^{13} \Delta I(t-i)\right) - \delta I(t) \qquad (3.13)$$

where $6w_1 + 7w_2$ equals 1. As a result of this equation, we get the findings in Figure 3.6.

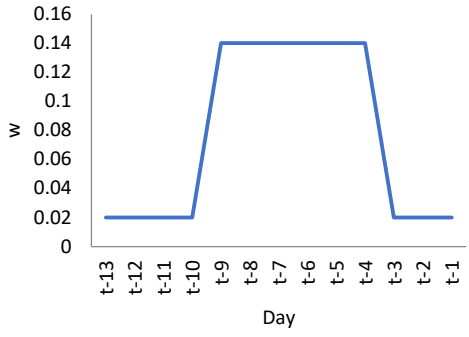
For the curves in Figure 3.6, the value of w_2 is smaller than the value of w_1.

Because w_2 is near to zero, we consider the values of w_2 to be noise and set w_2 to zero. Finally, (3.9) can be rewritten as follows:

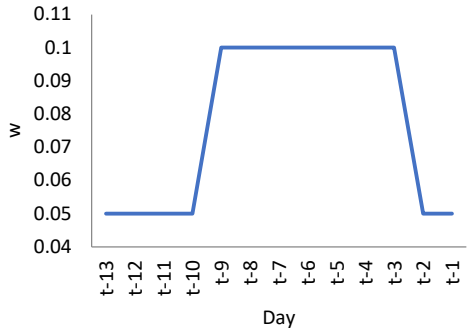
$$\frac{dI(t)}{dt} = \beta_4(t) \sum_{i=4}^{9} \Delta I(t-i) - \delta I(t) \qquad (3.14)$$



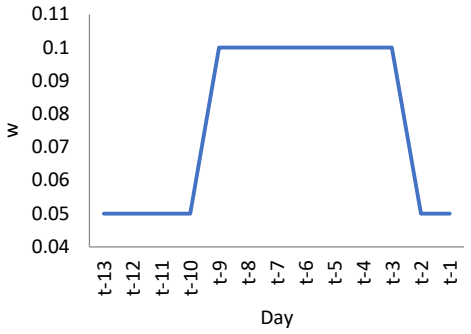
(a)



(b)



(c)



(d)

Figure 3.6 Effect of Newly Infected Cases in Days (t – 13) to (t – 1) on Infected Cases on Day t in(a)Maharashtra, (b)Kerala, (c)Karnataka, and (d)Delhi

3.3.2. COVID-19 Cases Prediction

The numbers of pre-processed infected cases from March 1-May 16, 2021, have been utilized as the training data to forecast the number of confirmed cases from May 17, 2021, till May 24, 2021.

In order to validate the efficiency of our model, the embedding of vaccination rate, and the effect of media publicity and public awareness about preventive measures, we perform a comparison between the conventional SIR, the SIQRV, the SIQRV with ES and LSTM, the SIQRV with ES, and LSTM embedded with PAN-LDA features. We compare the daily forecast, Mean Absolute Percentage Error (MAPE), and the determination coefficient, R^2 (R-Square), for Maharashtra, Kerala, Karnataka, and Delhi.

$$R^2 = \left(\frac{\sum_{i=1}^{N}(a_i-\bar{a})(f_i-\bar{f})}{\sqrt{\sum_{i=1}^{N}(a_i-\bar{a})^2}\sqrt{\sum_{i=1}^{N}(f_i-\bar{f})^2}} \right)^2 \tag{3.15}$$

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{a_i-f_i}{a_i}\right| \tag{3.16}$$

where a_i represents the observed values and the f_i are the corresponding forecasted values on the i^{th} day respectively. $\bar{a}$ and $\bar{f}$ denotes the mean value of actual and forecasted values. N represents the total number of forecast days.

The comparative results are given in Tables 3.1-3.4.

Table 3.1 Comparison of the Predicted and Actual COVID-19 Confirmed Cases in Maharashtra

Date	SIR	SIQRV	SIQRV+ES +LSTM	SIQRV+ES+LSTM +PAN-LDA	Actuals
17 May	31691889	31192712	**31358437**	31370280	31338407
18 May	32976321	32973786	32509423	**32267081**	31588717
19 May	31246499	31292862	31339409	**31461952**	31874364
20 May	27086291	27322825	31836470	**31913034**	32154275
21 May	26908162	27260526	31614879	**32427652**	32441776
22 May	33798249	33711704	33703059	**33360525**	32723361
23 May	**33036852**	33071176	33481204	33227730	33013516
24 May	35209522	36737138	36776979	**33301504**	33277290
R^2	0.3250	0.4089	0.7850	**0.8812**	
MAPE	5.9039	6.6878	2.8901	**0.8757**	

Table 3.2 Comparison of the Predicted and Actual COVID-19 Confirmed Cases in Kerala

Date	SIR	SIQRV	SIQRV+ES +LSTM	SIQRV+ES+LSTM +PAN-LDA	Actuals
17 May	18538374	18461298	17773601	**18129076**	18014842
18 May	18172927	18215038	**18172927**	18503720	18149395
19 May	**18268519**	18165616	18525175	18615175	18289940
20 May	18000224	18968922	19674717	**18444988**	18421465
21 May	18313600	18340709	18511600	**18578558**	18555023
22 May	**18657837**	18439810	18646598	18448909	18681051
23 May	19368790	18551914	**18717931**	19119721	18794256
24 May	19540265	19305119	18916822	**18905911**	18881587
R^2	0.7198	0.5511	0.5371	**0.7785**	
MAPE	1.6762	1.5588	1.3211	**0.9652**	

Table 3.3 Comparison of the Predicted and Actual COVID-19 Confirmed Cases in Karnataka

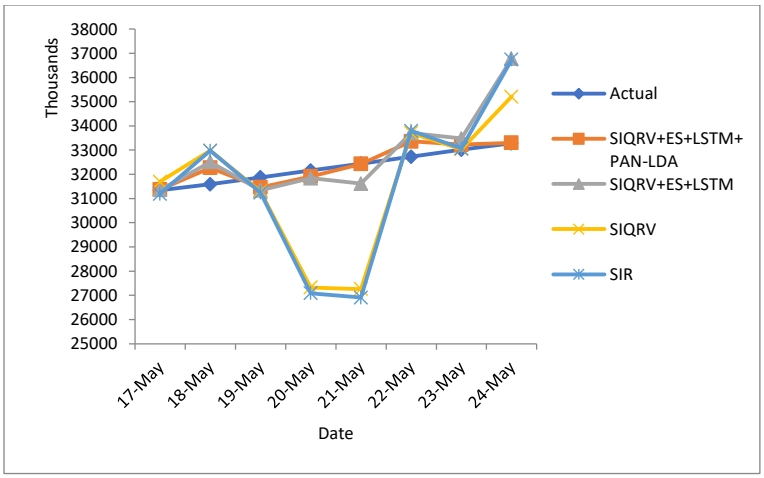
Date	SIR	SIQRV	SIQRV+ES +LSTM	SIQRV+ES+LSTM +PAN-LDA	Actuals
17 May	27744501	**28064531**	27655612	**28014400**	27976933
18 May	28057856	**28074546**	28035938	28127727	28070180
19 May	27764393	28222841	**28205152**	28075484	28199718
20 May	28373754	28086105	28855063	**28315084**	28320429

21 May	28020899	**28488088**	28411099	28399989	28453442
22 May	28235857	28346878	28457969	**28546770**	28582203
23 May	28360876	29171965	**28713756**	28453866	28707320
24 May	**28772799**	28879507	28862406	28702822	28816043
R^2	0.8205	0.8556	0.8275	**0.9542**	
MAPE	0.8369	0.5027	0.4930	**0.2982**	

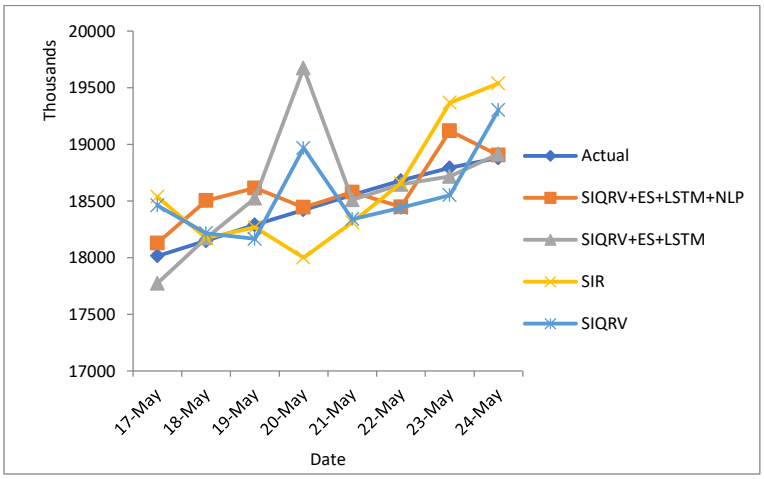
Table 3.4 Comparison of the Predicted and Actual COVID-19 Confirmed Cases in Delhi

Date	**SIR**	**SIQRV**	**SIQRV+ES +LSTM**	**SIQRV+ES+LSTM +PAN-LDA**	**Actuals**
17 May	18294693	18017137	18717330	**18389871**	18342482
18 May	**18412220**	18732831	18446435	18483929	18407486
19 May	18478783	18350347	18477307	**18472433**	18474059
20 May	18409091	18160472	19411746	**18571550**	18532803
21 May	17922381	18604425	**18600825**	18522320	18595993
22 May	19007880	18531836	**18663985**	18686022	18659148
23 May	18694718	18735939	18771023	**18724819**	18727191
24 May	19163520	18867071	18715085	**18736310**	18788697
R^2	0.6309	0.6799	0.1271	**0.9535**	
MAPE	1.0798	0.9264	0.9616	**0.2154**	

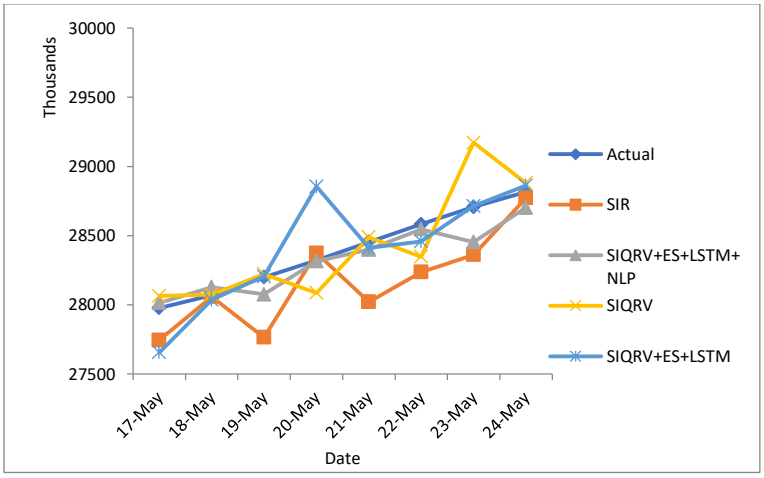
As shown in Figure 3.7, our model produces reasonable predictions. Our proposed model is a significant advancement over the traditional SIR approach. The ES and LSTM network does not consistently improve compared to the SIR model, making it unstable. The Deep-SIQRV model predicts more precisely in comparison to other models.


Thousands
38000
37000
36000
35000
34000
33000
32000
31000
30000
29000
28000
27000
26000
25000
17-May
18-May
19-May
20-May
21-May
22-May
23-May
24-May
Date
Actual
SIQRV+ES+LSTM+ PAN-LDA
SIQRV+ES+LSTM
SIQRV
SIR

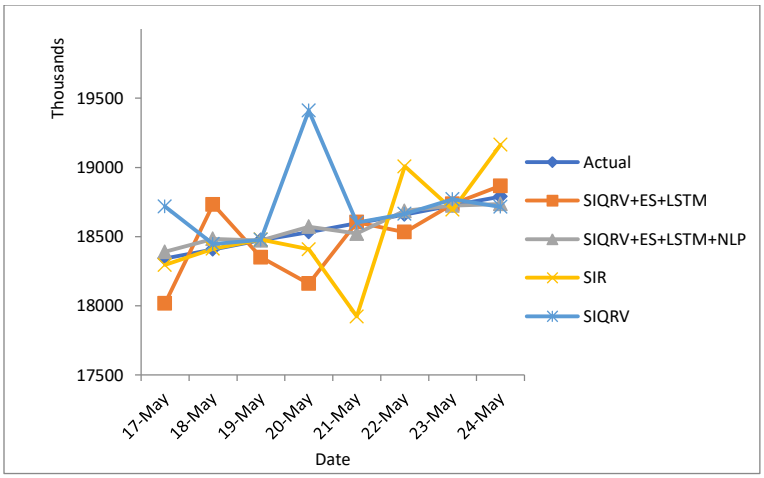
(a)


Thousands
20000
19500
19000
18500
18000
17500
17000
17-May
18-May
19-May
20-May
21-May
22-May
23-May
24-May
Date
Actual
SIQRV+ES+LSTM+NLP
SIQRV+ES+LSTM
SIR
SIQRV

(b)


Thousands
30000
29500
29000
28500
28000
27500
17-May
18-May
19-May
20-May
21-May
22-May
23-May
24-May
Date
Actual
SIR
SIQRV+ES+LSTM+ NLP
SIQRV
SIQRV+ES+LSTM

(c)



(d)

Figure 3.7 Actual and Predicted Confirmed Cases for(a) Maharashtra(b) Kerala(c) Karnataka(d) Delhi

3.3.3. Result Analysis

First, we analyzed the relationship between the new confirmed cases in the 13 days, and then new daily confirmed cases on day t in Maharashtra, Kerala, Karnataka, and Delhi. Similar to the conclusions in Section 3.3.1, the new confirmed cases from the day $(t-9)$ to $(t-4)$ have a larger contribution to the new confirmed cases on day t, whereas the contribution rates of the new confirmed cases from the day $t-13$ to $t-10$ and $t-3$ to $t-1$ are lower as shown in Figure 3.6. Therefore, the average infection time is around 5.5 days.

The performance metric indicated in Section 3.3.2 is used to assess the performance of the following models: SIR, the SIQRV, the SIQRV with ES and LSTM, the SIQRV with ES, and LSTM embedded with PAN-LDA features. Each of the above-stated models was trained with the collected data set and predicted coronavirus confirmed cases. Based on the above analysis, we estimated the number of infected cases from May 17, 2022, to May 24, 2022. And the corresponding daily confirmed case is used to verify the model. Tables 3.1-3.4 give quantitative results of different models for different cities. On using the PAN-LDA model, the proposed model achieved a high R^2 value, which further illustrates the effectiveness of the PAN-LDA model. Taking Karnataka as an example, the R^2 value of daily production is increased to 0.9542 compared with only considering the SIQRV model.

For the daily forecast, our model predicted amounts of daily cases across Maharashtra, Kerala, Karnataka, and Delhi with relatively high accuracy in which all the values of R^2 are greater than 0.77, especially for cities like Karnataka and Delhi values of R^2 of daily prediction achieve 0.9542 and 0.9535.

Also, Figure 3.7 shows the pandemic trends for Maharashtra, Kerala, Karnataka, and Delhi from May 17 2022, to May 24 2022. From the figure, we can find that there is a fine synchronization between the number of predicted cases and that of ground truth cases for daily prediction. It visually proves the prediction ability of our model in the pandemic. The proposed model can be served as a reference for the decision to support epidemic prevention and control of novel diseases. This result demonstrates that NLP features give additional information and guidance for disease prediction. To summarise, in this chapter, an AI model for forecasting the coronavirus cases based on the Deep-SIQRV model is proposed, which includes a PAN-LDA module that incorporates valuable information from the news articles and posts from the social media platform Reddit from the public into the forecast calculation process, resulting in more accurate forecast results.

3.4. Chapter Summary

In this chapter, a Deep-SIQRV, an epidemiological model based on machine learning and deep learning methodologies, was proposed to perform COVID-19 trend predictions, finding out that infection rates of reported cases vary at various time intervals. The model focuses on the consequences of waning immunity caused by the conversion of vaccinated individuals back to susceptible ones. This proposed model used the LSTM network and ES to optimize the infection rate and other parameters of the model. This article employed the PAN-LDA model to analyze and extract COVID-19-related news items and social media posts and encoded them as semantic features. The infection rate of the proposed Deep-SIQRV model is then updated using these features, which were embedded into the LSTM network. According to this article, infected cases at different time intervals contribute differently to susceptible individuals, which seeks to anticipate the trend of COVID-19. The effect of confirmed cases from several days prior to day t on newly confirmed cases at day t is investigated. We suggest a grouped multiparameter technique, which divides the infection rates of recent cases into separate time-based groups. The forecasted results of the model are found to be extremely consistent with the empirical data, representing that the proposed model was more effective in accurately assessing the development trend and transmission law of the virus. Also, incorporating language information such as related news and social media posts enhances the accuracy of the prediction model. This article shows that the efficiency with which data are realized is critical for creating a novel epidemic prevention system.

Chapter 4 EFFECT OF NON-MEDICAL HEALTH DETERMINANTS FOR PANDEMIC PREDICTION

Given the uncertain nature of the coronavirus disease during the early stages of the pandemic, it is challenging to settle for a specific factor related to the spread of the disease. This chapter geospatially studied the social mobility, demographic, health, socio-economic, and climatic parameters connected with the COVID-19 distribution. Though it is not clear what influence these factors have on COVID-19 transmission, this chapter, taking India as a study area during unlock-phases of the first wave of COVID-19, analyzed 1) the spatial and temporal patterns in coronavirus monthly growth rate, 2) the factors associated with the COVID-19 distribution and how these factors affected the COVID-19 growth rate, and 3) the interaction effects of the different variables on the Coronavirus growth rate, based on GeoDetector. Later, we investigate the impact of associated multi-source variables, such as social mobility patterns, climatic indicators, and air pollution, on the COVID-19 cases during the first phase of the second wave in order to forecast the possibility of the third wave. Following that, we proposed a short-term fixed-effect multiple regression model based on the multi-source data to forecast daily confirmed cases. Section 4.1 intends to explore how environmental, health, socio-economic and demographic variables impact the evolution of the COVID-19 pandemic. Section 4.2 presents a short-term prediction model using multi-source data during the initial phases of the coronavirus waves. The comparative analysis of the proposed model with traditional methods is also discussed. Section 4.3 wraps the summary of the research work.

4.1. COVID-19 Growth Rate Correlation with Influencing Factors

Human coronavirus transmission is frequently associated with human movement, climatic variables, such as temperature and humidity, environmental factors, etc. The MERS and SARS-related Coronavirus, for example, have both been linked to cold weather [222]. However, whether SARS-CoV-2 transmission is susceptible to non-medical determinants, in the same way remains unclear.

Given the lack of effective drugs and vaccines, as well as the virus's highly infectious nature, many nations have implemented Non-Pharmaceutical Interventions (NPIs), such as international and domestic travel restrictions, stay-at-home orders, cancellation of public events and gatherings, and school and workplace closures in an attempt to slow the dissemination of coronavirus and flatten the epidemic curve. In order to deal with a contagious and infectious disease, it is challenging to settle for a specific factor related to the spread of the disease. Government-implemented NPIs do not seem to justify the COVID-19 growth rate. The previous study findings indicate that the demographic, climatic, and social variables,

such as hospital beds, life expectancy, pollution, Global Health Security detection index, temperature, and urban agglomeration, played a significant role in explaining the growth rate [223]. A key question in the coronavirus pandemic is why the epidemic growth rate varies too much across regions. These variations imply that the COVID-19 growth rate is affected by various factors. In addition to the various public health approaches to the pandemic variations in the outcome of curve flattening may be clarified by taking into account several demographic, social, and climatic parameters, including, but not limited to, population aging [224], country development, income and Gross Domestic Product (GDP) [225], pollution [226], temperature [227], and population density [228,229] among others.

Our objective is to study the influence of environmental variables on the coronavirus growth rate during the unlock phase of the first wave of COVID-19 in India. This study also intends to explore how health, socio-economic and demographic variables impact the evolution of pandemic, which could have an enormous impact on virus growth and be further examined on global and local levels. This research work is not intended to be conclusive; rather, it intends to help expose epidemiological base variations throughout various states in India, form the direction of further research on possible future contagious diseases, and understand their transmission in general.

Table 4.1 presents the climatic, economic, demographic, and health variables included. Later, these factors were incorporated into the proposed model to investigate their impact on prediction performance.

Table 4.1 Climatic, Demographic, Socio-Economic, and Health Factors

Variable type	Predictors
Climatic	> Temperature(^{0}C) > Humidity (mm)
Demographic	> Population aged 60 and above (%) > Population density (per sq. km) > Urban Population (Urbanization)
Socio-economic	> GDP per capita (Cr INR) > Poverty rate
Health	> Life expectancy (in years) > Hospital beds available

4.1.1. Material and Methods

4.1.1.1. Data Collection

The COVID-19 dataset for India was gathered from the PRS Legislative Research dataset [230] for June 1, 2020, and November 31, 2020.

Each state and Union Terrority's humidity and temperature data were obtained from TuTiempo.net [231], which offers the researchers a significant volume of worldwide weather data. The State-wise population density data were collected from the 2011 census of India, which was available online at https://www.census2011.co.in [232]. The GDP per capita data and elderly population data are collected from the official website of the Ministry of Statistics and Program Implementation, Government of India [233]. The life expectancy, poverty rate, and urban population data are gathered from [234]. The data on hospital bed availability is accessible from the official website of The Center For Disease Dynamics, Economics & Policy [235].

4.1.1.2. Logistic Growth Model

Because of its simple concept and rapid calculation, logistics is frequently employed in regression fitting of time series data. LotKa [17] developed the application of the logistic growth model. Beyond their biological foundations, logical functions are now used in a time series prediction challenge, such as epidemic modeling. Logistic growth is defined by increasing growth in the early stages but diminishing growth as you approach the maximum. Individuals did not take stringent precautions at the start of the outbreak, and the number of infected individuals was limited at first; as a result, the number of infected sources gradually increased. When the infection base attained a certain level, the pandemic situation exhibited an enormous exponential growth pattern, and then, with public support and government intervention, the pandemic situation progressively slowed the dissemination speed, eventually reaching the maximum number of cumulatively infected individuals. The inflection point represents the time when the daily number of cases reaches a maximum, marking the turning point at which disease spread begins to fall. As long as the data contain this critical point and the time period immediately after that, the number of cases and curve-fitting forecast would be pretty accurate. The development of a logistic growth model is advantageous because it can explain the rise and decrease in cases around the inflection point. Suppose the logistic curve converges at an initial phase; in that case, it indicates that the epidemic is under control, implying that the pandemic-control measures are effective. If, on the other hand, the logistic curve continues to expand exponentially even beyond the inflection point, it suggests that the measures place to control the epidemic are ineffective, and the pandemic is still in its initial stages.

For example, in the case of coronavirus disease, logistic growth is characterized by exponential growth at an early stage and eventually slows, approaching a limit at the end of the outbreak, i.e., the maximum of

infections. The COVID-19 growth rate during the unlock phase for each state in India was determined by a logistic growth model [236]. The rate of change in total reported coronavirus infected cases was fitted using a logistic growth model to the number of infected cases per day. In order to do this, the least-square fitting technique was used. In mathematical epidemiology, when using the logistic model [237], the following equation can be used to define the growth of the number of cases over time :

$$\frac{dc}{dt} = rc\left[1 - \frac{c}{K}\right] \tag{4.1}$$

where c is the total number of infection cases over time t, r is the disease growth rate, and K indicates the final pandemic size

The solution of (4.1) is:

$$c = \frac{K}{1+\left[\frac{K-c_0}{c_0}\right]exp^{-rt}} \tag{4.2}$$

where c_0 denotes the number of cases at the initial stage.

The change in total cases determined using a seven-day rolling average of infected cases is as follows:

$$I_t = c_{t+t} - c_t \tag{4.3}$$

where t denotes a small change in time.

The change in total confirmed cases opted instead of total confirmed coronavirus cases since the observations derived from the same cumulative curve are associated. Both the curves, the total cases and the change in total cases, hold the same information from a mathematical viewpoint, but from a statistical point of view, they do not. The assumption that the errors in observations of most curve fitting algorithms are statistically independent is invalid for cumulative curves, where all cases from previous observations are included in the current observation [223].

4.1.1.3. Spatial Autocorrelation Analysis

The spatial autocorrelation analysis determined how a unit's value has a spatial correlation with the surrounding unit's values [238]. Because the statistics relied exclusively on studies independent of one another, spatial autocorrelation was evident. If autocorrelation occurred in a map, it ruled out the possibility of the studies being independent of one another. In order to assess global spatial correlation, the Global Moran's Index (GMI) was utilized [239], and the Local Indicators of Spatial Association's Local Moran's Index (LMI) was utilized to assess local spatial correlation [240]. The spatial autocorrelation study was performed using the ArcGIS software's spatial statistics and pattern analyzing tool. We ran the spatial autocorrelation analysis in ArcGIS using the default parameters.

$$Global\ Moran's\ I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(X_i-\bar{X})(X_j-\bar{X})}{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}\sum_{i=1}^{n}(X_i-\bar{X})^2} \quad (4.4)$$

$$Local\ Moran's\ I\ = \frac{n(X_i-\bar{X})\sum_{j=1}^{n} W_{ij}(X_j-\bar{X})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} \quad (4.5)$$

where W_{ij} is the spatial weight between feature i and j,$(X_i - \bar{X})$ is the deviation of an attribute for feature i from its mean, and n is equal to the number of features. Global Moran's I have a value between [-1,1]. If the index is near zero, the distribution exhibits a random pattern. The distribution is scattered if the index is near -1 and clustered if the index is near 1. The LISA aggregation can categorize the growth rate distribution into two types of positive correlations based on Local Moran's I calculations: Low-Low (LL) and High-High (HH) correlations; two forms of negative correlations: Low-High(LH) and High-Low(HL) correlations; and non-significant correlations. Positive correlations suggest that the growth rate is consistent with the observed trend in the surrounding region. The HH type denotes locations with high values with high-value neighbors. In contrast, the LL type denotes locations with low values with low-value neighbors. Negative correlations suggest that the growth rate is inversely related to the surrounding region's trend. The HL type denotes locations with high values with low-value neighbors. In contrast, the LH type denotes locations with low values with high-value neighbors. Non-significant correlations indicate that the growth rate follows a random pattern [241].

4.1.1.4. Multiple Linear Regression Analysis

Linear regression is a primary and commonly used predictive analytic technique. The primary goal of regression is to examine how accurately a set of predictor variables predict a dependent variable, which specific factors are key determinants of the dependent variable, and how they impact it? These regression estimations are used to describe the relationship between one or more independent variables and one dependent variable. The model is represented in Eq. (1), where y and x are the dependent and independent variables, β_0 is the intercept, and β_1 is the regression parameter as slope and $\in$ is the random error.

$$y = \beta_0 + \beta_1 x + \in \quad (4.6)$$

Linear regression has limitations in that it frequently investigates a relationship between the mean of the input variables and the mean of the output variables. Just as the mean does not completely explain a single variable, LR does not deliver a clear understanding of variable correlations. As a result, the Multiple Linear Regression (MLR) model is employed to analyze the numerous components. In this scenario, the dependent variable (goal variable) relies on a large number of independent factors. A regression equation with numerous variables may be written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_p x_{ip} + \in \quad (4.7)$$

where

y_i= target variable, where i=1,….,n observations,

x_{ij} = value of the pth variable X_p for the i^{th}case,

β_0 is the y-intercept,

β_j , $j=1,2,....p$, is the slope coefficients with respect to variable X_p , and

$\in$ is the error term of the model

The chapter focuses on determining each daily active case or new confirmed COVID-19 case using a regression model that would aid in estimating the country's future situation while taking non-medical health variables into account.

Later, to detect multicollinearity, we employ a metric known as the Variance Inflation Factor (VIF), which measures the correlation between one predictor and the other predictors in a model. Interestingly, it was found that the VIF value for humidity was very high. Consequently, this factor was removed from the model for further research.

4.1.1.5. Geodetector method

The Geodetector approach may quantify and assess the importance of stratified spatial heterogeneity [242]. GeoDetector's sub-detectors are factor detector and interaction detector. The q value is used to express factor detection [243], i.e. the relative importance of explanatory variables, and its formulations are as follows:

$$q = 1 - \sum_{h=1}^{L} N_h \sigma_h^2 / N\sigma^2 = 1 - SSW/SST \quad (4.8)$$

$$SSW = \sum_{h=1}^{L} N_h \sigma_h^2 \quad (4.9)$$

$$SST = N\sigma^2 \quad (4.10)$$

where

N_h and σ^2 stands for the number of units and variance of observations within the $h^{th}(h = 1, 2 \ldots, L)$ study region

N and σ_h^2 denotes the number of units and variance of the observations in the total study region

SSW denotes the weighted sum of local variance

SST represents the weighted sum of global variance.

The interaction detection can determine whether the independent variables can improve the explanatory power of the dependent variable when they work together, or their effect on the dependent variable is independent of each other. It can identify whether there is an interaction between the two variables, how

strong the interaction is, and whether the connection is linear or non-linear. Compared to traditional statistical approaches, the Interaction detector has a significant benefit in distinguishing the q values, i.e. ($q(X_1)$ and $q(X_2)$) to the interaction q value, i.e. ($q(X_1 \cap X_2)$). The q value represents the capacity of interactions of two independent variables to explain the dependent variable [244].

4.1.2. Results: Influence of Factors on COVID-19 Transmission

This section discusses the experimental results of the spatial and temporal spread and patterns in coronavirus monthly growth rate. Also, we presented the results of the spatial analysis of the exploratory data along with the factors associated with the COVID-19 distribution. We also discussed how these factors affected the COVID-19 growth rate and their interaction effects on the pandemic growth rate based on GeoDetector.

4.1.2.1. Spatial and Temporal Spread of COVID-19 Growth Rate

As shown in Table 4.2, the nationwide average growth rate decreased from June to November 2020. In June, the average growth rate was 0.113, and in November, it was 0.024. The highest level of growth rate between June and November 2020 was among the eastern states of India. The growth rate's coefficient of variation varied from 123.751% to 71.525%. After June 2020, the differences in growth rate among different states were small, with the coefficient of variation below 72%.

Table 4.2 COVID-19 growth rate descriptive statistics from June to November 2020

	Jun	**Jul**	**Aug**	**Sept**	**Oct**	**Nov**
Min	0.016	0.011	0.007	0.018	0.008	0.0009
State/UT	Chandigarh	Ladakh	Delhi	Manipur	Mizoram	Dadra and Nagar Haveli and Daman and Diu
Max	0.787	0.145	0.132	0.085	0.085	0.058
State/UT	Sikkim	Meghalaya	Nagaland	Chandigarh	Meghalaya	Arunachal Pradesh
Mean	0.113	0.053	0.047	0.044	0.042	0.024
Standard Deviation	0.140	0.026	0.0242	0.017	0.018	0.017
Coefficient of Variation	123.751%	50.346%	51.324%	39.203%	42.696%	71.525%

Figure 4.1 represents the high level of growth rate that was primarily distributed in eastern states in June 2020, while in July, the growth rate was also observed to increase in southern states such as Andhra

Pradesh, Telangana, and Karnataka. There was a growth rate increase in central states from September to November.

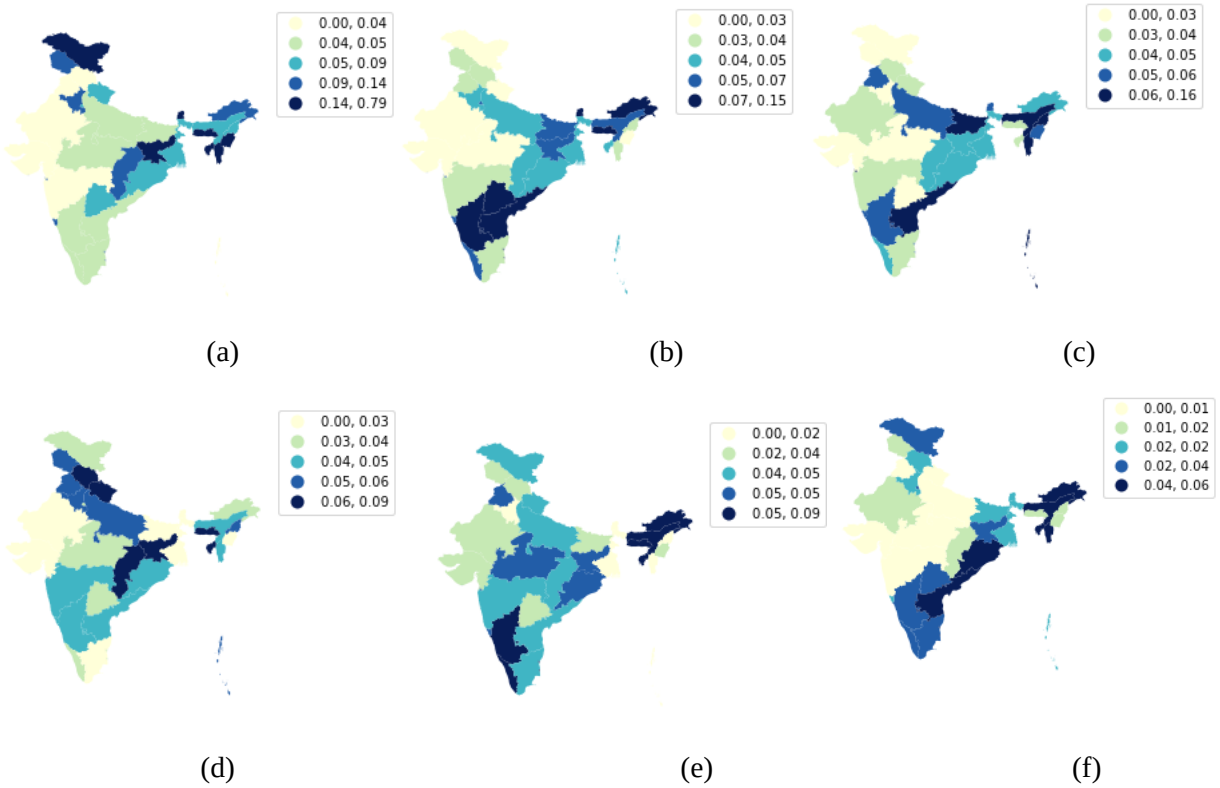


Figure 4.1 Spatial Variation of COVID-19 Growth Rate from June to November 2020

4.1.2.2. Spatial Trend of COVID-19 Growth Rate

The growth rate was significantly clustered in all the studied months. The Global Moran's I varied from 0.15 to 0.51($P < 0.05$) i.e. June (0.18), July (0.19), August (0.51), September (0.148), October (0.165) and November (0.24). Furthermore, spatial aggregation decreased between August and September. High GMI was found in August (0.51), while the lowest was registered in September (0.15).

The COVID-19 growth rate spatial pattern changed as the virus spread (Figure 4.2). The LMI findings revealed that the growth rate spatial distribution was mostly composed of Low-Low clustering types. From June to August and November, the Low-Low cluster types were predominantly witnessed in most of central and northern India. Low-High clusters were discovered in eastern India from June to August and November. High-Low clustering types were mostly detected in northern regions with a sparse distribution. High-High clustering types were found irregularly throughout different parts of the country.

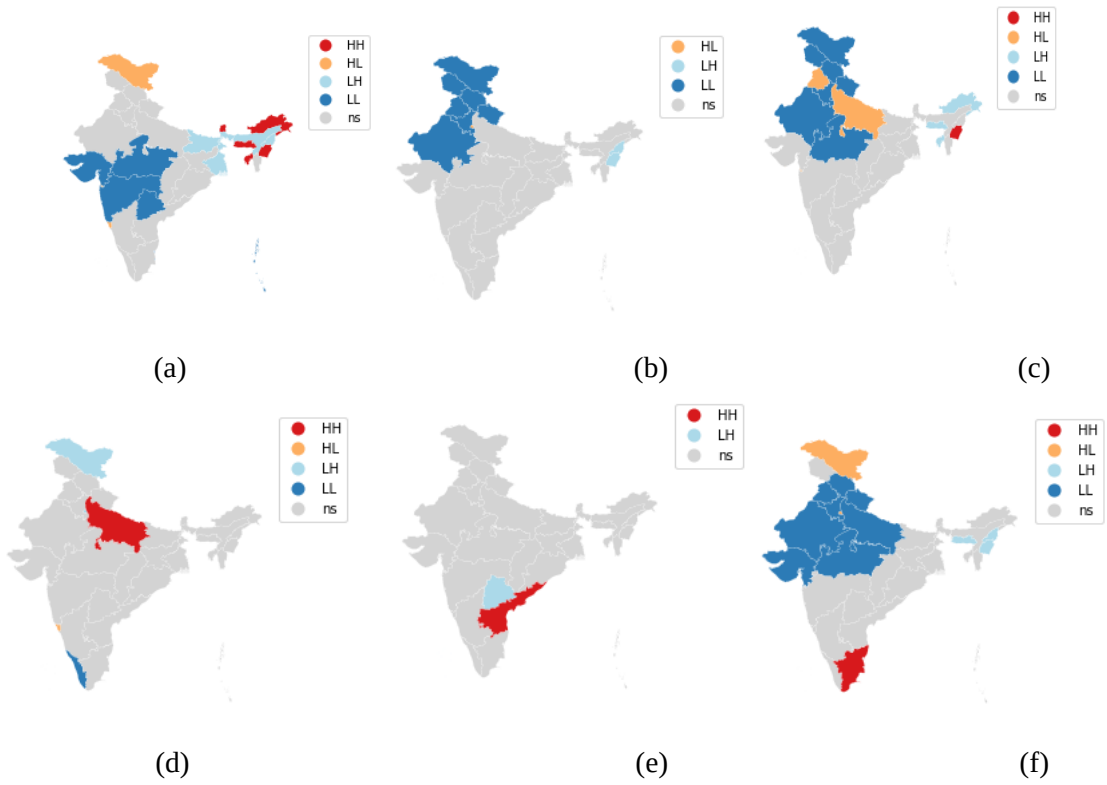
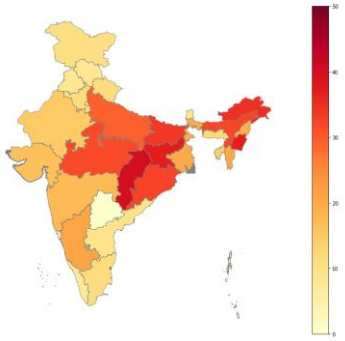


Figure 4.2 Local Aggregation of COVID-19 Growth Rate from June to November 2020

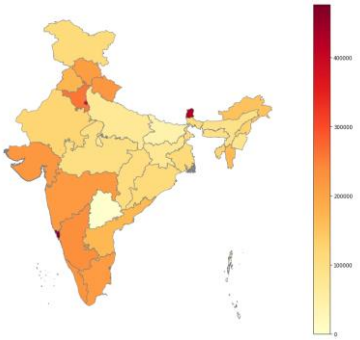
4.1.2.3. Spatial Analysis of Exploratory Data

Figure 4.3 displays the spatial variation of the demographic, socio-economic, and health variables mentioned in Table 4.1. It is essential to mention that there was missing information on various variables for certain states/UTs. For instance, no official data is available for the poverty rate and percentage of the older population for Ladakh and Telangana. In addition, the research and outcomes of this work are based solely on the available data.

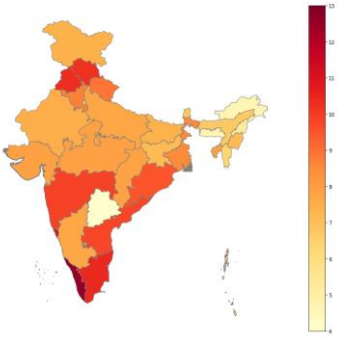
As seen in the figure, the poverty rate is highest in central and eastern parts of the country, especially Chhattisgarh, Jharkhand, Manipur, Arunachal Pradesh, etc. At the same time, GDP is higher in Goa and Sikkim, and southern India, mainly in Karnataka and Kerala. The percentage of older adults(aged 60 or above) is much more regulated, with the maximum count in northern and southern parts of the country, i.e., Tamil Nadu, Kerala, Punjab, Himachal Pradesh, etc. The figure shows the spatial variation of the urban population distribution across various regions in India. The A&N, Sikkim, and D&N hold the minimum percentage of the urban population. The other focus of the pandemic in India was Delhi, the most densely populated state in the country, as seen in the figure. The last two figures show the state-wise distribution of hospital beds and life expectancy.



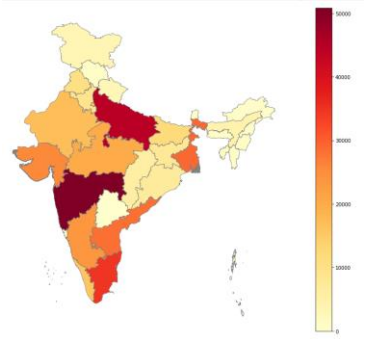
Poverty Rate



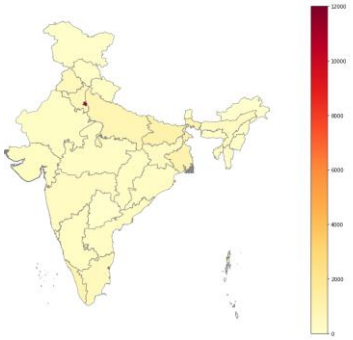
GDP



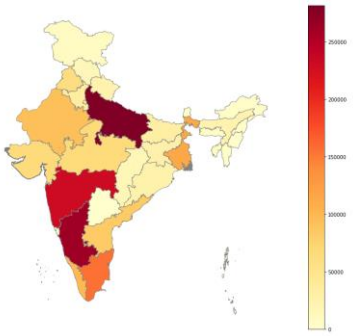
Older Population (%)



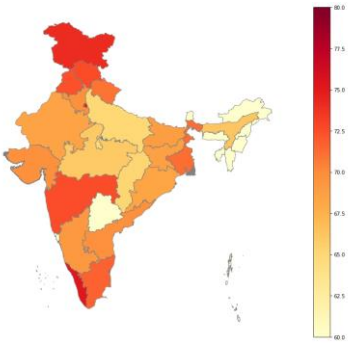
Urban Population



Population Density



Number of Hospitals Beds



Life Expectancy

Figure 4.3 The Spatial Distribution of Variables

4.1.2.4. Analysis of Factor detection

All the variables' q values cleared the significance test at the 5% level, showing that these variables have a considerable capacity to predict the spatial distribution of the coronavirus pandemic spread. Precisely, the $q(p)$ values of Temperature, Population aged 60 and above, Population density, Urban Population, GDP per capita, Poverty rate, Life expectancy, Hospital beds available were equal to 0.021(0.003), 0.15(0.000), 0.17(0.000), 0.01(0.002), 0.08(0.000) ,0.04(0.001), 0.01(0.006) and 0.072(0.001) respectively. Nevertheless, population density, population aged 60 and above, GDP, and hospital beds available had a comparatively high impact and can explain 16.86%, 14.92%, 7.90%, and 7.20% of the variance, respectively.

4.1.2.5. Analysis of Interaction Detection

The interaction detector reflects the influence of the interaction between two independent variables on the COVID-19 growth rate.

Table 4.3 q Value of Variables Interaction Effect on COVID-19 Growth Rate

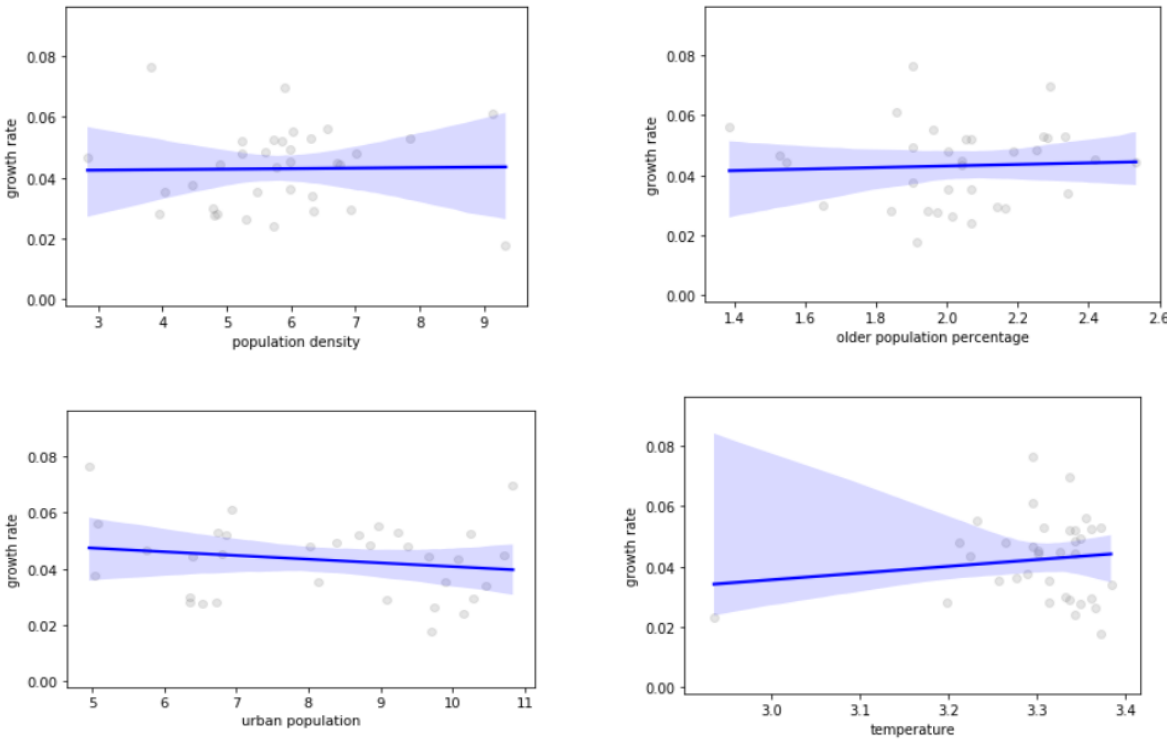
	T	**age>=60**	**PD**	**UB**	**GDP**	**PR**	**LE**	**HB**
T	0.021							
age	0.19	0.15						
PD	0.23	0.45	0.17					
UB	0.05	0.19	0.22	0.01				
GDP	0.12	0.3	0.32	0.19	0.08			
PR	0.12	0.20	0.22	0.07	0.16	0.04		
LE	0.03	0.19	0.2	0.03	0.12	0.07	0.1	
HB	0.14	0.22	0.25	0.14	0.24	0.16	0.12	0.072

*T, age>=60, PD, UB, GDP, PR, LE, and HB represents Temperature, Population aged 60 and above, Population Density, Urban Population, GDP per capita, Poverty Rate, Life Expectancy, Hospital Beds available, respectively.

Table 4.3 shows the q value of interaction was greater in Population density ∩ elder population (0.45), Population density ∩ GDP (0.32), GDP ∩ elder population (0.3), Population density ∩ Hospital beds available (0.25), and GDP ∩ Hospital Beds (0.24).

4.1.2.6. Correlation Analysis of Exploratory Data and COVID-19 Growth Rate

After performing VIF analysis, the humidity covariate was removed from the multiple linear regression model. Figure 4.4 and Table 4.4 report the results from data fitting using the multiple linear regression model to the covariates. The adjusted R-squared and multiple R-squared values for the model were found to be 0.731 and 0.791. These statistics show that the model has an excellent explanatory capacity. The percentage of the older population, population density, poverty rate, and humidity positively correlate with the growth rate. On the other hand, the Urban Population, Life expectancy, total hospital beds, GDP, and temperature negatively correlate to the growth rate. Table 4.4 shows that, among all the covariates, GDP(p=0.001) was the most significant variable, followed by the percentage of the older population(p=0.011), average temperature(p=0.018), and life expectancy(p=0.038).



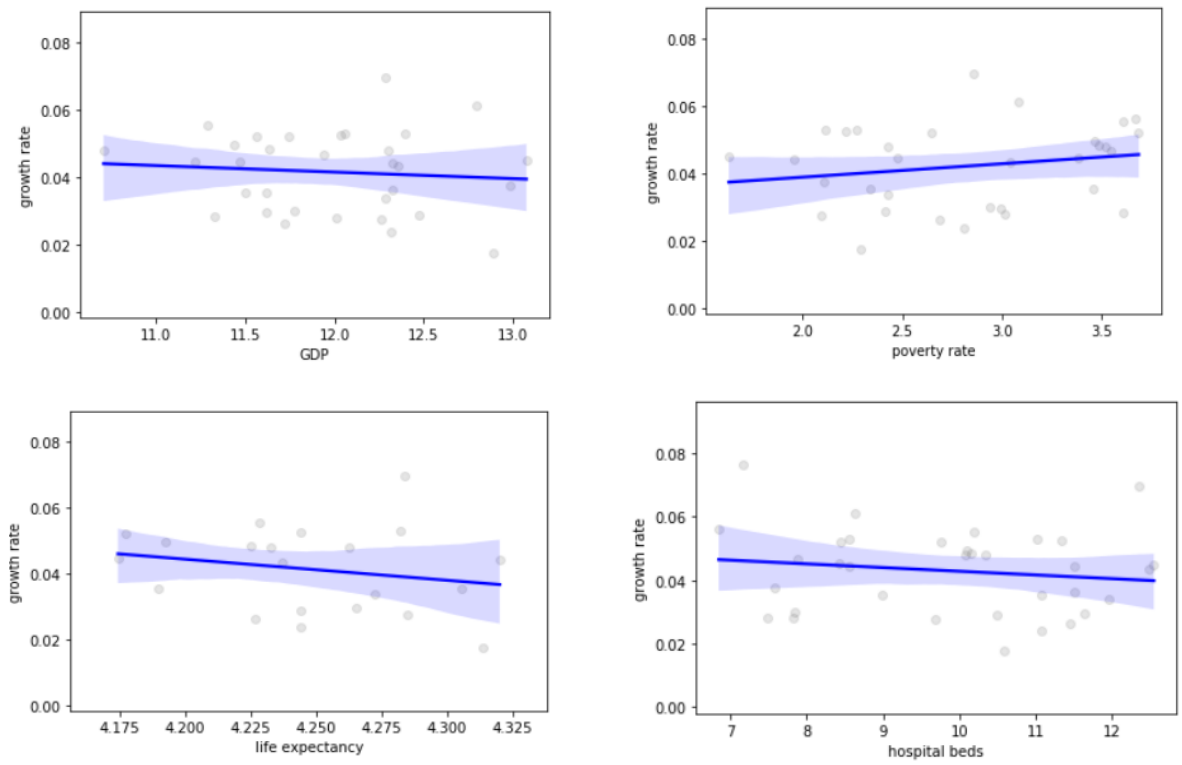


Figure 4.4 The Relationship between Independent Variables and COVID-19 Growth Rate

Table 4.4 Statistical Summary of Multiple Regression

Variables	**coef**	**std err**	**t**	**P>|t|**
Population Density	0.532	0.0262	2.028	0.052
Older Percentage	0.0078	0.003	2.729	0.011
Urban Population	-0.0828	0.0923	-0.897	0.377
Temperature	-0.0020	0.001	-2.503	0.018
GDP	-0.0219	0.0058	-3.789	0.001
Poverty Rate	0.0008	0.000	1.674	0.105
Hospital Beds	-0.0167	0.0153	-1.095	0.283
Life expectancy	-0.0005	0.000	-2.182	0.038

4.1.3. Discussion

The current study has overcome the limitation of earlier univariate analyses that did not consider the interdependence between various factors. Based on the collected data, the results indicate that the most

significant variables affecting the growth rate were the GDP, percentage of the older population, average temperature, and life expectancy. When considering the correlation of factors with the epidemic growth rate, GDP, the percentage of the older population, average temperature, population density, and life expectancy have shown the most significant relationship. While population density, urban population, poverty rate, and hospital beds were found to be insignificant.

The population density has a positive but not particularly significant association with the coronavirus growth rate. This is especially true for Delhi, with the maximum density, but the growth rate is not high (Figure 4.4). The CDC corroborates that respiratory droplets are one of the primary sources of propagation of coronavirus disease, supporting the role of population density in disseminating the disease [245]. Maintaining a safe distance from other people becomes more difficult in densely populated areas. Since coronavirus spreads by human interaction [246], it is widely assumed that coronavirus propagates quickly in dense regions, although the risk of infection is less in low-population density areas. This result is consistent with numerous studies that found a modest correlation between population density and coronavirus disease [228,247]. However, on the contrary, a recent investigation found that the infection rate is inversely [248] or unrelated to population density [249].

The percentage of the older population positively correlates with the coronavirus growth rate. It has been found that aged adults have higher infection and mortality rates; younger adults, on the other hand, have milder symptoms and often asymptomatic conditions [250]. It is interesting to note that Maharashtra, having the highest older population [233], has suffered the maximum COVID-19 cases and a high growth rate. It is worth stating that older adults are more susceptible, and their case and death rates are among the largest of all age groups. A recent meta-analysis of over half a million coronavirus cases from various countries revealed the impact of age on mortality [251]. This is consistent with the findings of this study, which positively associate the older population and the COVID-19 growth rate (Figure 4.4).

Next, it was found that the urban population is negatively associated with the growth rate. According to Spatio-temporal modeling studies, coronavirus propagated rapidly from urban to rural areas and did not remain limited to urbanized environments [252]. According to one study [253], deaths from infectious diseases in urban counties decline significantly faster than in rural areas. A possible explanation is that the population in metropolitan areas has access to excellent urban healthcare facilities favorable for COVID-19 patients' treatment. Also, people in rural areas are often less likely to seek medical services or indulge in preventive health practices than urban residents.

In this research work, the temperature is negatively correlated with the coronavirus growth rate. Delhi and Rajasthan were among the states that observed the highest average temperature [231] and have recorded lower COVID-19 growth rates (Figure 4.1 and Figure 4.3). This result contrasts with the existing literature, which indicates that average and maximum temperature are significantly and positively

associated with coronavirus incidences [254]. In contrast, another study shows that warm climates may be able to prevent or at least delay the propagation of coronavirus disease, despite the fact that this variable can only describe only a small fraction of the variability in the viral distribution [255]. The findings of this study reveal that temperature has a close association with SARS-CoV-2 risk, and the chance of coronavirus transmission will reduce as the temperature increases.

In the present work, both Life expectancy and GDP substantially impacted the COVID-19 growth rate and were negatively correlated with the pandemic growth rate. Compared to Life Expectancy, GDP has a more significant impact on coronavirus growth rates. On the other hand, the poverty rate positively correlates with the coronavirus growth rate. One can expect that states having a high poverty rate are likely to have a higher growth rate of coronavirus disease. One reason is that regions with a high poverty rate are expected to have insufficient access to better medical and healthcare facilities, resulting in the fast propagation of the disease. It shows that COVID-19 is less prevalent in affluent regions.

The number of available hospital beds is negatively associated with the growth rate. It can be expected that states having higher numbers of hospital beds are likely to have faced less stress and burden, necessitating less severe and restrictive interventions. Instead, some studies showed that having more hospital beds and several other healthcare resources available will help the fast propagation of COVID-19. This could reduce treatment standards, resulting in insufficient access to ventilators and a rise in the coronavirus mortality ratio [256]. This is consistent with the research [257] undertaken in the United States, which found that increased COVID-19 deaths were linked to a rise in the beds in non-intensive care units. However, the results of this study reveal a particularly negative association between hospital beds and COVID-19 cases (Figure 4.4), implying that intensive medical care is required to slow COVID-19 growth. It is noticed that the eastern states of India have the less number of hospital beds and are among the regions with the maximum COVID-19 growth rate (Figure 4.3 and Table 4.2).

4.1.3.1. Limitations

Lastly, it is necessary to discuss some limitations of this research. Firstly, it is well known that there is significant underreporting of COVID-19 cases due to limited testing. Therefore, the confirmed coronavirus-infected cases may not completely reflect the real state of the pandemic, which may impact the accuracy of the results. As long as there is a constancy in the underreporting of coronavirus cases across the study area, the systemic geographic bias was not considered in reporting. However, it is worth mentioning that the number of actual cases would also certainly be higher. Second, the sample size for the covariates is also inadequate. As a result, some variables' β-coefficients were low, causing their correlations with the coronavirus risk to be speculative. Also, data for some covariates are not available for certain states/UTs. Lastly, the assumption of a constant growth rate is a major limitation, despite the

fact that the coronavirus growth rate varies over time and can be influenced by numerous epidemiological, socio-economic, and health factors, causing a change in the growth rate over time for different regions.

As the geographic spread of the emergent infectious disease can cause rapid multipoint transmission of the outbreak, it is an effective approach to analyze the epidemic situation base on the multisource data at the initial stages of the pandemic. To showcase the impact of multisource data in the early phases of the epidemic, we focused on the human mobility and environmental data during the initial stages of the second wave of coronavirus and proposed a prediction model.

4.2. Pandemic Prediction at Early Stages

The coronavirus pandemic is returning in different waves. Due to the uncertain nature and limited knowledge of the transmission characteristics of the coronavirus at the beginning of the pandemic, there was not a clear understanding of the factors responsible for the spread of the disease and its new variants. Therefore, it is crucial to analyze the factors at the initial stages of the pandemic and the trend of the upcoming waves of the pandemic. In order to address this issue, this chapter also explores the related multi-source data at the initial stages of the COVID-19 wave and presents a short-term prediction model by including the correlated determinants.

Before the first wave of cases could erupt, the central government promptly implemented a nationwide lockdown, limitations on overseas travel, and a slew of other stringent measures (COVID appropriate behavior). Therefore, even after the implication applied by the government in the nation and a consistent drop in the virus during the first in December 2020, the second wave of COVID-19 appeared in mid-February. However, it is believed that the second wave of coronavirus, which appeared in mid-February 2021 in India, was fueled by people lowering their guard and participating in social activities, as well as contradictory messages from the government which allowed political rallies and religious gatherings.

During the COVID-19 second wave in India, there was a drop in coronavirus cases after having the highest number of cases on May 6, 2021. A detailed examination of the pattern of reported cases in India in recent weeks indicates that the COVID-19 second wave has not yet touched its base. Since July 1, 2021, India has registered around 40,000 to 50,000 new cases (monthly average range), accounting for 0.12 percent to 0.15 percent of the overall cases. On September 1, 2021, the country reported more than 47 thousand new coronavirus cases, with five states, i.e., Kerala (32,803), Maharashtra (4,456), Tamil Nadu (1,509), Andhra Pradesh (1,186), and Karnataka (1,159) accounting for more than 88% of the total cases. Kerala and Maharashtra have seen an increase in R_o, having a threshold of more than 1. Many alarming findings suggest that if proper measures are not adopted quickly, the disease may spread out of control in these states [258].

The second wave has demonstrated the critical need to concentrate on pandemic preparation. As a result, it is critical to examine the state of the epidemic in the early phases of the several waves of the coronavirus outbreak using multi-source data. However, little knowledge about the transmission characteristics of the new variants of the disease can be obtained during its early stages. Moreover, the pandemic could be complex and fluctuate over time and space; it is challenging to accurately predict the transmission dynamics parameters. Especially, tracking the social mobility effects on reducing coronavirus transmission remains unclear.

4.2.1. Multisource Data Explored at the Early Phase of COVID-19 Wave

This chapter focuses on the daily data from February 20, 2021, to March 27, 2021, corresponding to the initial phase of the coronavirus second wave in India [259]. Figure 4.5 presents the time plot of the confirmed cases. Given the present transmission potential and mode of the COVID-19 pandemic, the potential influence of social mobility cannot be neglected [260]. Given this, we concentrate on the two explanatory factors. The first variable is movement data in public areas from February 20, 2021, to March 12, 2021, which is accessible in Google's Community Mobility Reports (CMR). This research depicts changes in mobility in six different types of locations: "retail and leisure," "workplaces," "parks," "transit stations," "grocery and pharmacy," and "residential" [261]. The percentage difference is determined by comparing current mobility in certain locations to average movement prior to lockdown. Public venues such as parks, workplaces, retail and recreation, and transport stations exhibited a negative percentage change over the research period, indicating a drop in mobility. Simultaneously, the positive percentage variation in mobility to supermarket and pharmacy sites suggests an increase in mobility in these areas [261]. The next variable is the environmental variable, including temperature, humidity, and pollution. The researchers received the average temperature and humidity data from the Weather Underground portal [262], which provided a large volume of global weather data. Furthermore, air quality data was obtained from https://aqicn.org [263].

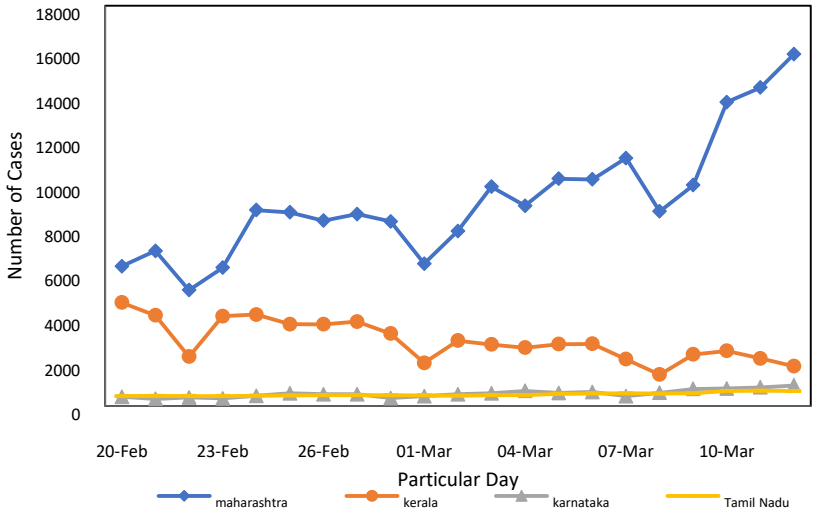


Figure 4.5 Time Plot of Daily Confirmed Cases

4.2.2. Coronavirus Growth Rate

The growth rate of a contagious disease in a region is a key indicator of how quickly it propagates and where it reaches its peak value. The growth rate of the contagious disease is usually calculated by using exponential, logistic, and Hierarchical logistic equations to fit the number of cases during the early stages of the disease [211,264,265]. The coronavirus disease dynamics are qualitatively close to those of a logistic curve since the number of total cases increases exponentially at first but then slows and reaches a peak. As a result, we use the logistic model and the least-squares fitting approach to approximate the growth rate of the COVID-19 second wave for each state in India [236]. A logistic growth model is utilized to fit the rate of change in the total number of confirmed coronavirus cases to the daily confirmed cases. We choose the logistic growth model to model the coronavirus pandemic as compartmental models, including the SIR models, have a complex parametrization process and are associated with a high level of uncertainty due to the limitation of control interventions, especially in the initial stages of the pandemic.

On the other hand, data-driven phenomenological models are not subject to these constraints.COVID-19 disease growth has been modeled using logistic models, employing generalized equations to account for various growth curves in different countries [266]. K. Wu et al. calibrated several phenomenological models to show the viability of using the logistic approach to model COVID-19 prediction [265]. Indeed, the model can provide reliable estimates of the thresholds of the coronavirus-related scenario.

4.2.3. Correlation Analysis

With respect to Maharashtra, we used correlation analysis to study the relationship between human mobility, air quality, relative humidity, average temperature, and COVID-19 cases. In India, the coronavirus second wave was reported in the mid of February 2021. Since the virus needs time to incubate and analyzing the specimen takes time, it is fair to conclude that it can spread to others during this period. According to the current literature, the average incubation time is between 5 to 12 days [267–269], after which some infected people show symptoms [264,270] while others remain asymptomatic [271]. As the prevalence of asymptomatic carriers is not accurately reported, they can become the sources of virus transmission unless some steps are taken to limit their mobility. As a result, we consider the incubation time to be 8-9 days prior to the onset of the disease. On a particular day t, in the association between the confirmed cases and the multi-source data, the number of cases may be correlated to the past lags of the latter. Since certain conditions can cause the transmission to lag, we extend the incubation period to 14 days.

In terms of various incubation periods, we investigate the relationship between the daily confirmed cases in Maharashtra from February 20, 2021, to March 12, 2021, and the associated daily google mobility data and other climatic influences. The Pearson correlation coefficient analysis was used to investigate the association between multi-source data and newly confirmed coronavirus cases. The following is a description of the modified Pearson correlation:

$$pearsonr = \frac{\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)}{\sqrt{\sum_{i=1}^{n}(x_i - m_x)^2}\sqrt{\sum_{i=1}^{T}(y_i - m_y)^2}} \tag{4.11}$$

where:

$x = [x_1, x_2, \ldots, x_n]$ and $y = [y_1, y_2, \ldots, y_n]$ represent the two vectors of length n, and m_x and m_y are the means of x and y, respectively.

4.2.4. Fixed-Effect Multiple Regression (FE_MR)

In this section, the proposed model generates the forecast of daily new coronavirus cases based on multi-source data and the historical confirmed cases. Unlike other conventional transmission dynamic models, our proposed model is data-driven. The historical confirmed cases and the factors that correlate with the daily confirmed cases are used as inputs without loss of generality.

We consider F as the set of these variables, where $f \in F$ represents a factor in F, such as temperature, human mobility, humidity, etc. Fixing the disease incubation period, the per day value of the factor f in T time interval is denoted as:

$$[f_{t-T+I}, \dots, f_t] \tag{4.12}$$

and the corresponding cumulative confirmed cases are denoted by the following sequence: $[y_{t+1}, \dots, y_{t+T}]$.

Based on the above historical observations, the below equation mathematically represents the proposed model:

$$y_{t+n} = \beta_0 + \sum_{f \in F} \beta_{f,i} f_{t-i} + \alpha_i y_{t-i-1} + W_{t+n} + u_{t-i}, \qquad i = 1,2, \dots\dots, T \tag{4.13}$$

where

y_{t+n} : number of confirmed cases on $(t+n)^{th}$ day and n is the time step length of the forecast,

β_0 is the intercept,

$\beta_{f,i}$ are the coefficients for the various factors,

α_i represents the time-invariant, fixed-effect component,

W_{t+n} is the weekend dummy; W_{t+n} has a value of 1 if t is a weekend and 0 otherwise, and

u_{t-i} is a time-varying error component are the model's trainable parameters.

To estimate the model's parameters, we used a supervised machine learning approach.

We applied the log () transformation for confirmed cases and the response variable to achieve strictly positive predicted values. Later, we back-transformed the mentioned transformation to produce forecasts.

We employed two distinct benchmark approaches to compare their prediction performance with the proposed model. The two most often used regression models for time series forecasting are Least Absolute Shrinkage And Selection Operator(LASSO) [272] and Ridge Regression(RR) [273]. In addition, LASSO and RR are multilinear models that can simultaneously manipulate more than one attribute [274]. These benchmark approaches are used to determine whether applying the proposed model adds value.

The correlation analysis is conducted using the Pearson correlation coefficient and the p value. On the next prediction task, we looked at variables having a high correlation and a p value less than 0.05.

4.3. Prediction Result using FE_MR Model

The forecasting metrics are the R^2, Root Mean Squared Error (RMSE), MAPE, and Mean Square Error (MAE). These matrices are calculated as follows:

$$R^2 = \left(\frac{\sum_{i=1}^{N} (a_i - \bar{a})(f_i - \bar{f})}{\sqrt{\sum_{i=1}^{N} (a_i - \bar{a})^2} \sqrt{\sum_{i=1}^{N} (f_i - \bar{f})^2}} \right)^2 \tag{4.14}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (a_i - f_i)^2} \tag{4.15}$$

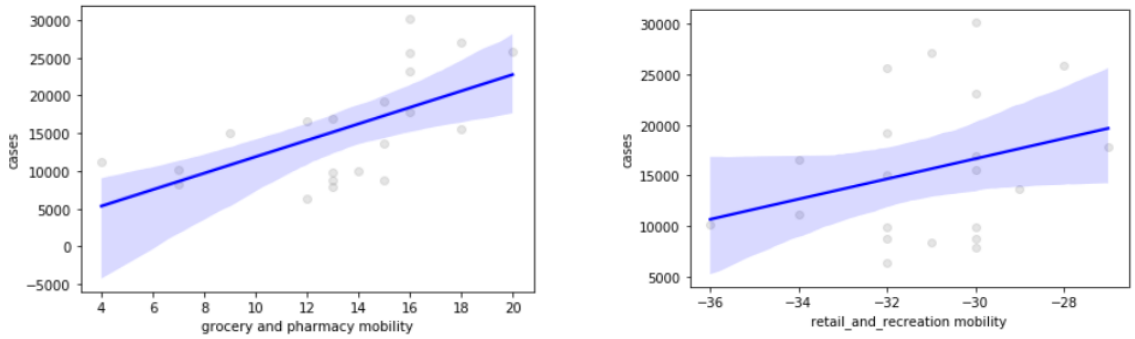
$$MAPE = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{a_i - f_i}{a_i} \right| \quad (4.16)$$

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |a_i - f_i| \quad (4.17)$$

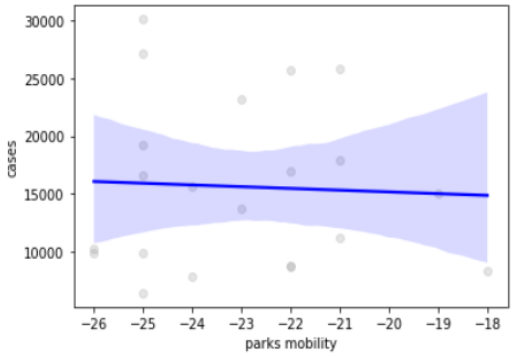
where a_i represents the observed values and the f_i are the corresponding forecasted values on the i^{th} day respectively. $\bar{a}$ and $\bar{f}$ denotes the mean value of actual and forecasted values. N represents the total number of forecast days.

These metrics help evaluate the forecast accuracy of the models from different perspectives. RMSE, MAPE, and MAE are concerned with the degree of deviation from the true value, whereas; R^2 shows the proportion of variance in the independent variable towards the dependent variable. A higher value of R^2 and the low once of RMSE, MAPE and MAE represent the estimation of great performance.

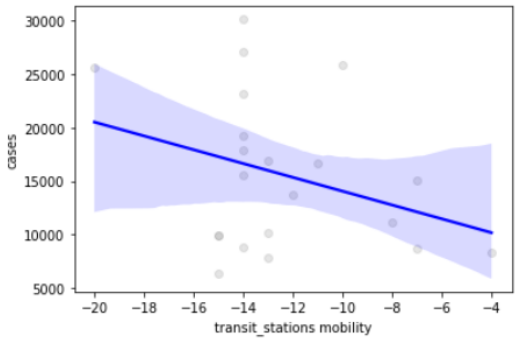
Following that, we examined the relationship between multi-source data and COVID-19 confirmed cases. Human mobility during the initial phases of the second wave of coronavirus plays a critical role in contagious pandemic transmission. A visualization of the Pearson correlation coefficient between the coronavirus cases and the observed variables with a 9-day lag is seen in Figure 4.6. First, it is very intuitive to notice from Figure 4.6(a) that there is a significant correlation between the grocery and pharmacy mobility trends and cases of coronavirus infections, with a 9-day lag showing the greatest significance (pearsonr= 0.755, p<0.00075). Meanwhile, the other mobility categories in Figure 4.6(b)-(e) show a weak positive correlation with the new cases. It implies that the social distancing restrictions implemented in public places might effectively control the COVID-19 transmission. Noticeably, according to Figure 4.6(f) and (g), the temperature shows a negative correlation with the daily confirmed cases (pearsonr=-0.112), while relative humidity is weakly positively correlated (pearsonr= 0.665, p value =0.001). Additionally, the results of air pollutants concentration are shown in Figure 4.6(h)-(j). The findings suggest that PM2.5 shows a weak positive correlation (pearsonr=0.364, p=0.104), whereas PM10 and NO_2 are negatively associated with reported outbreak cases (pearsonr = -0.006 and -0.285, respectively).



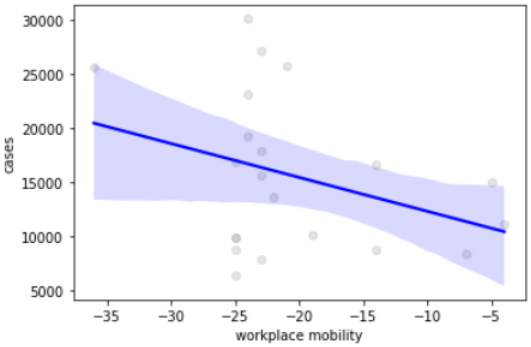
(a)


cases
parks mobility

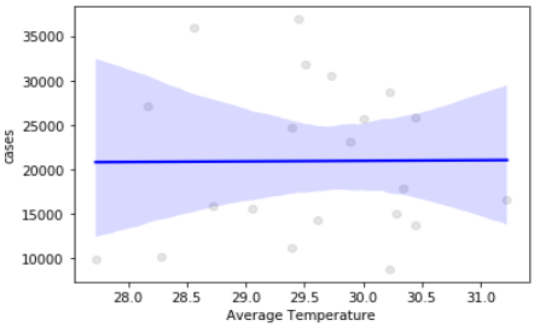
(b)


cases
transit_stations mobility

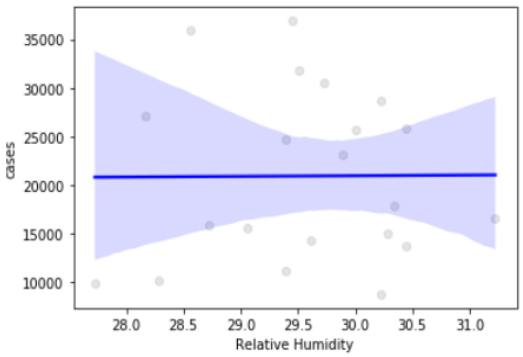
(c)


cases
workplace mobility

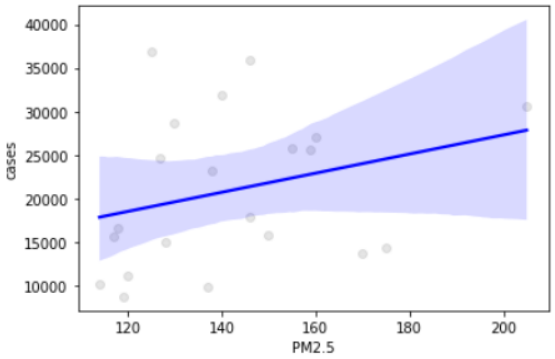
(d)


cases
Average Temperature

(e)

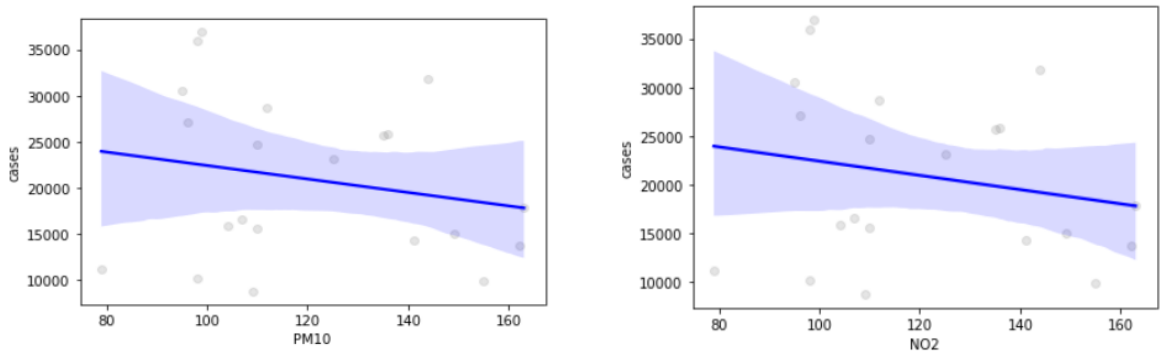

cases
Relative Humidity

(f)


cases
PM2.5

(g)

(h)



(i) (j)

Figure 4.6 Correlation Analysis between Daily Confirmed Cases and Multi-Source Data from February 20, 2021, to March 12, 2021, which consists of (a) Grocery and Pharmacy Mobility, (b) Retail and Recreation Mobility, (c) Parks Mobility, (d) Transit and Stations Mobility, (e) Workplaces Mobility, (f) Average Temperature, (g) Relative Humidity, (h) PM2.5, (i) PM10 and (j) NO_2

Table 4.5 summarizes the correlation coefficient between the observed factors and confirmed cases for different lag days. In order to validate if any variables have a lag effect, we have extended the interval to 14 days.

Table 4.5 The correlation coefficient between the observed factors and confirmed case

Variables		**n=6**	**n=7**	**n=8**	**n=9**	**n=10**	**n=11**	**n=12**	**n=13**	**n=14**
GPM*	Pearsonr	0.546	0.581	0.704	0.755	0.573	0.582	0.5247	0.6125	0.5719
	p value	0.0010	0.005	0.0003	0.0007	0.0055	0.0056	0.0014	0.0031	0.0057
RR*	Pearsonr	0.128	0.093	0.234	0.379	0.283	0.283	0.050	0.192	0.080
	p value	0.580	0.688	0.306	0.090	0.213	0.213	0.831	0.404	0.729
Parks Mobility	Pearsonr	0.145	0.084	-0.126	0.104	0.178	0.240	0.124	0.161	0.067
	p value	0.532	0.718	0.588	0.653	0.441	0.294	0.592	0.487	0.774
Transit and Station Mobility	Pearsonr	-0.084	-0.092	-0.321	-0.299	-0.250	-0.036	-0.199	-0.115	-0.108
	p value	0.719	0.691	0.156	0.188	0.274	0.875	0.387	0.620	0.641

Workplace Mobility	Pearsonr	0.120	0.143	-0.099	0.155	-0.081	0.096	0.022	0.077	0.110
	p value	0.606	0.536	0.671	0.502	0.728	0.679	0.926	0.741	0.634
PM2.5	Pearsonr	0.148	0.305	0.313	0.364	0.296	0.147	0.120	0.161	0.325
	p value	0.522	0.179	0.167	0.104	0.192	0.524	0.604	0.485	0.150
PM10	Pearsonr	-0.284	-0.224	-0.005	-0.006	0.021	-0.049	-0.220	-0.240	-0.203
	p value	0.212	0.330	0.984	0.980	0.929	0.832	0.337	0.295	0.378
No_2	Pearsonr	-0.513	-0.478	-0.342	-0.285	-0.426	-0.436	-0.474	-0.433	-0.414
	p-value	0.017	0.028	0.129	0.211	0.054	0.048	0.030	0.050	0.062
Temperature	Pearsonr	-0.097	-0.022	-0.112	0.105	0.061	0.090	-0.032	-0.042	0.015
	p value	0.674	0.924	0.629	0.650	0.792	0.697	0.891	0.858	0.949
Humidity	Pearsonr	0.627	0.550	0.665	0.504	0.452	0.472	0.515	0.585	0.489
	p-value	0.002	0.009	0.001	0.020	0.040	0.031	0.017	0.005	0.025

*GPM: Grocery and Pharmacy Mobility; RRM: Retail and Recreation Mobility

The statistical findings suggest that grocery and pharmacy mobility indicates the highest consistency, while other mobility categories have a negative or weak consistency. The results indicate that, during the initial phases of the second wave of Coronavirus in India, among all the mobility categories, grocery and pharmacy mobility plays a critical role in the number of daily COVID-19 cases. Moreover, it can be observed that the p values for the grocery and pharmacy mobility for every case are less than 0.05. It should be noted that grocery and pharmacy mobility has the clearest association with the daily confirmed cases at a 9-day lag effect, which corresponds with the incubation time discussed in the literature. Based on the above observation, we further perform the experiments using grocery and pharmacy mobility to verify its importance in forecasting coronavirus cases.

As per the correlation analysis discussed in the above section, the effect of grocery and pharmacy mobility on early dissemination is chosen as the key factor for the short-term prediction of coronavirus cases. Figure 4.7 illustrates the time plot of the grocery and pharmacy mobility data from February 20, 2021, to March 12, 2021. Even though the time series plot is noisy, yet some systematic trends can be easily noticed.

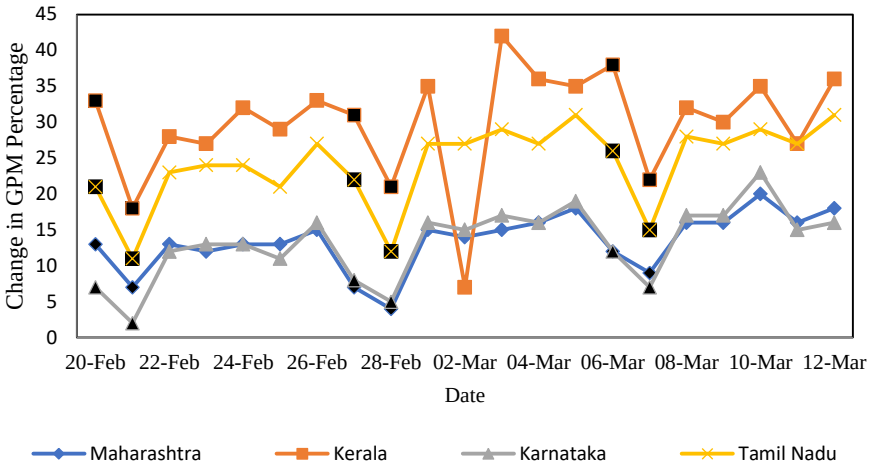


Figure 4.7 Time Plot of Percentage Change in Mobility from a Baseline to Grocery and Pharmacy Places, where Marker Fill with Black Denotes the Change in GPM Percentage on Weekends

The time plot shows a pattern that the percentage change in the "grocery and pharmacy" mobility is lower on the weekend than the working days. Firstly, we only use grocery and pharmacy mobility as an input to our proposed model to further illustrate its efficacy during the initial phases of the second wave. In order to forecast the number of infected cases for Maharashtra, Kerala, Karnataka, and Tamil Nadu, the period was set from March 13, 2021, to March 27, 2021. Table 4.6 demonstrates the overall performance of forecasting models using R^2, RMSE, MAPE, and MAE.

Table 4.6 One Day Ahead Prediction Performance with Single Factor in the Different States of India

	Maharashtra			**Kerala**			**Karnataka**			**Tamil Nadu**		
Accuracy Measure	**RR**	**LASSO**	**FE_MR**	**RR**	**LASSO**	**FE_MR**	**RR**	**LASSO**	**FE_MR**	**RR**	**LASSO**	**FE_MR**
R^2	0.532	0.503	**0.671**	0.555	**0.576**	0.471	0.684	0.640	**0.857**	0.763	0.742	**0.918**
RMSE	3387.48	3561.93	**2954.36**	767.986	734.589	**448.425**	592.217	632.821	**535.047**	271.906	273.039	**194.599**
MAPE	1.108%	1.110%	**1.107**%	1.174%	1.197%	**1.147**%	1.212%	1.384%	**1.207**%	1.230%	1.215%	**1.052**%
MAE	2078.87	2207.53	**1680.8**	653.666	613.333	**390.933**	488.466	456.933	**421.2**	220.928	224.714	**176.071**

As shown in Table 4.6, using only grocery and pharmacy mobility data, compared to LASSO and RR, the highest R^2 value is the highest achieved by our model for all the states, i.e., Maharashtra (0.671), Kerala

(0.471), and Karnataka (0.857) and Tamil Nadu (0.918). Table 4.6 further shows that while grocery and pharmacy mobility can depict the overall trend, it is unable to portray changes in reported cases effectively. As a result, employing only one factor limits the model's forecasting efficiency. Thus, to even further increase the forecasting accuracy, we included the historical case data. As a result, the R^2 value of the projection has increased to 0.789, 0.513, 0.915, and 0.961 for Maharashtra, Kerala, Karnataka, and Tamil Nadu, respectively (see Table 4.7), compared to only considering grocery and pharmacy mobility. Thus, it can be concluded that the proposed model clearly outperforms benchmark models for all forecast accuracy measures.

Table 4.7 One Day Ahead Prediction Performance with Multiple Factors in the Different States of India

	Maharashtra			**Kerala**			**Karnataka**			**Tamil Nadu**		
Accuracy Measure	**RR**	**LASSO**	**FE_MR**	**RR**	**LASSO**	**FE_MR**	**RR**	**LASSO**	**FE_MR**	**RR**	**LASSO**	**FE_MR**
R^2	0.650	0.630	**0.842**	0.637	0.643	**0.713**	0.721	0.718	**0.915**	0.829	0.802	**0.961**
RMSE	843.742	825.034	**770.596**	713.771	556.123	**416.024**	727.296	701.897	**420.386**	236.990	238.492	**164.313**
MAPE	1.371%	1.592%	**1.253**%	1.289%	1.293%	**1.272**%	1.141%	1.155%	**1.213**%	1.165%	1.152%	**1.179**%
MAE	718.467	710.6	**604.333**	553.6	429.866	**326.928**	637.214	607.214	**346.357**	203.285	206.857	**139.285**

In addition to the overall performance, we have also forecasted the results for the 1-to-5-day ahead. Table 4.8 reveals the Maharashtra, Kerala, Karnataka, and Tamil Nadu forecast results from 1-to5-day ahead. Our model forecasted the confirmed cases in the above states with a high degree of precision; for some states, the R^2 values are greater than 0.9. Particularly for Karnataka and Tamil Nadu, the R^2 values of daily prediction are 0.915 and 0.961, respectively, implying that the severity of the epidemic is very well linked to the amount of human mobility in these areas. This is supported by case counts from other states in India. However, the overall trend can be captured; the R^2 values started to decrease as the number of predicted days n increased. The following are the explanations for this phenomenon:

- As the number of prediction days,n, increases, there is a decrease in the number of training data for learning the model.
- The disparity in everyday mobility is still visible due to the close connection between "grocery and pharmacy" mobility and case records.

As a result, the absence of real-time data supplementation would impact the prediction model's accuracy.

Figure 4.8 depicts the trends in the pandemic for Maharashtra, Kerala, Karnataka, and Tamil Nadu from March 13, 2021, to March 27, 2021. It can be easily noticed from the figure that there is a close association between the daily predicted cases and the actual number of cases, which demonstrates our model's ability to forecast the early stages of a pandemic. At the beginning phases of the second wave of Coronavirus in Maharashtra, our proposed model acted as a guide for disease and pandemic prediction and control.

Table 4.8 n-Days Ahead Prediction of the Total Infected Cases by the Proposed Model for Maharashtra, Kerala, Karnataka, and Tamil Nadu

Region	Accuracy Measure	n=2	n=3	n=4	n=5
Maharashtra	**R^2**	0.791	0.754	0.622	0.496
	RMSE	9113.093	10047.713	12666.566	16558.677
	MAPE	38.54%	40.24%	38%	55%
	MAE	6308.733	6544.467	7653	9871.667
Kerala	**R^2**	0.577	0.555	0.541	0.428
	RMSE	583.540	423.650	474.220	1076.430
	MAPE	27.45%	16.44%	24.43%	58.36%
	MAE	246.62	318.82	321.21	528.46
Karnataka	**R^2**	0.901	0.905	0.826	0.501
	RMSE	478.320	693.440	361.830	761.820
	MAPE	23.76%	39.72%	14.73%	32.81%
	MAE	322.71	313.71	145.3	430.74
Tamil Nadu	**R^2**	0.956	0.868	0.772	0.510
	RMSE	177.870	341.470	431.270	612.870
	MAPE	17.33%	30.13%	25.59%	33.62%
	MAE	114.93	237.12	278.15	473.85

Appendix A

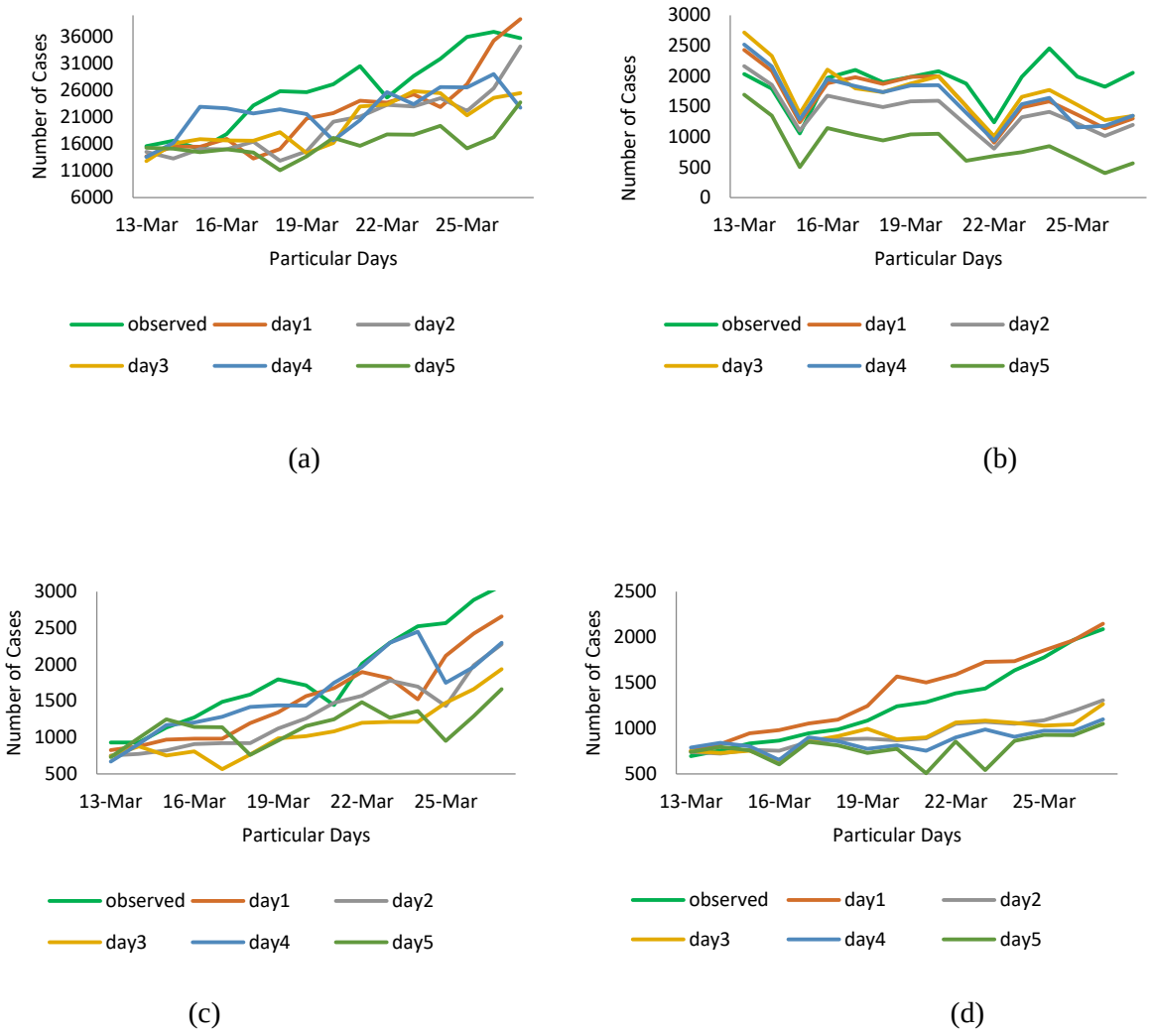


Figure 4.8 Number of Estimated Cases for (a) Maharashtra, (b) Kerala, (c) Karnataka, and (d) Tamil Nadu

Numerous challenges constrain our understanding of the coronavirus epidemic, including gaps in the environmental factors and demographic features of pandemic cases. Likewise, nations have adopted apparently comparable yet considerably different control policies, making it challenging to appreciate which approach is effective. In this chapter, we used publicly accessible datasets to show that, during the initial phase of COVID-19 second wave of the coronavirus, the dissemination of coronavirus cases in India can be explained by population mobility, especially grocery and pharmacy mobility. Looking at the trends of google mobility [261], there has been an increase in grocery and pharmacy mobility since August 2021, indicating a high possibility that India will face the third wave of coronavirus if no preventive measures are taken. From the healthcare disaster experience during the second wave, India

should emphasize the significance of different strategies for third-wave readiness. Also, early and massive responses from epidemic-affected regions are needed, which would significantly delay the spread of coronavirus by implementing initiatives such as lockdowns, social distancing, restrictions in grocery regions, and so on. One major limitation of this study is that we only chose to include mobility patterns and meteorological indicators, and the impact of vaccination on COVID-19 was not studied due to the low vaccine coverage and lack of data at the initial stages of the second wave in India, which was increased later and may impact the magnitude and timings of the future waves in the country.

Statistical models are essential for assessing infectious disease data analyses in real-time. During the initial stages of the different coronavirus waves in India, the proposed model can perform a short-term and accurate forecast of the epidemic distribution based on human mobility data. Simultaneously, it demonstrates the importance of adopting strategies such as imposing lockdown to prevent the rapid and large-scale dissemination of pandemics.

Furthermore, by carefully assessing the multidimensional variables such as climatic data, human behavior, and so on, a regional transmission risk model can be developed to study the development of contagious disease dynamics within the community, as well as more scientific and realistic recommendations for everyday protection of citizens and decision-makers.

4.4. Chapter Summary

This study presented a geospatial analysis of demographic, health, socio-economic, and climatic factors that influence the COVID-19 incidence in India. The study spans a 6-months timeframe that begins the day after the country proclaimed the unlock phase. Based on the available data, the findings of the current study support the hypothesis that climatic, demographic, health and socio-economic factors play a significant role in virus growth. Concerning climatic factors, a negative association was detected between coronavirus growth rate and temperature, whereas a positive association was found between coronavirus growth rate and humidity. Furthermore, an inverse association between COVID-19 growth rate was found with urban population and GDP per capita; in contrast, population density, percentage of older adults, and poverty percentage showed a positive association. Moreover, the health factors such as life expectancy and the availability of beds in hospitals have negative associations with the growth rate of coronavirus. The GDP, percentage of the older population, average temperature, and life expectancy were considerably correlated with the coronavirus growth rate in India. Also, the interactions between the dependent variables affected the pandemic growth rate. The interaction between the Population density and elderly population, GDP and Population density, and GDP and the elderly population had a more significant impact on the growth rate than each component alone. Based on a comprehensive assessment of non-

meteorological indicators, the policymakers can reform policies to keep in view the health protection and safety measures, demographics of the population, etc., across different states to prevent and mitigate coronavirus spread.

Later in this chapter, we use multi-source data to evaluate the association between daily confirmed cases and certain potential parameters during the early stages of the second coronavirus wave in India. We discovered that during the early stages of the coronavirus wave, the grocery and pharmacy movement had the strongest connection with the coronavirus pandemic spread. Based on mobility data and COVID-19 cases, we proposed a fixed-effect multiple regression model for estimating the coronavirus pandemic in the short term. The model will provide robust solutions for policymakers and decision-makers by tracking human motion in real-time. Consequently, they will be able to predict the possibility of future COVID-19 waves and develop necessary measures in time to save as many lives as possible. The influence of vaccination on the spread of COVID-19 will be included in future research.

Chapter 5 VECTOR REPRESENTATION OF DOCUMENTS USING FEATURE CLUSTERS

Various document clustering approaches have been developed throughout the years to group the textual data. Its major purpose is to categorize samples based on degrees of similarity. The efficiency with which text representations are used substantially influences the performance of document clustering algorithms. Feature extraction is the process of extracting a significant subset of features from a dataset to enhance the document clustering task. This chapter introduced a word cluster-based modified tf-idf (WC_MTI) method to improve the feature representation and enhance the document clustering. Furthermore, if we want to design an unsupervised framework, autoencoder is an interesting choice since it can map raw data into feature space by unsupervised recreating the data itself. Reconstruction loss is utilized initially to assist the autoencoder in learning the data representation, and then the K-means algorithm is implemented. Section 5.1 reveals the overview of the entire chapter. Section 5.2 briefly introduces existing methodologies employed in this chapter, such as Kernel Principal Component Analysis (KPCA), word2vec, Vector Space Model (VSM), K-means clustering, and Deep Embedded Clustering. We then discuss the proposed method, its algorithm for cluster-based document representation, and its details in Section 5.3. Following that, in Section 5.4, we discuss the experiments and results. Finally, section 5.5 concludes the summary of the entire chapter.

5.1. Overview

With the increasing growth of online text content, document clustering has received great interest from industry and academics. The document clustering method groups comparable texts into clusters using statistical data on the frequency of words, phrases, or sentences. The document clustering methods group the documents based on some feature matrix to ensure that the documents within a cluster are closer than those from other clusters [275]. To cluster textual data, the data must first be properly translated to a vector space, i.e., each text document must be vectorized before grouping the vectors using clustering methods such as K-means [276]. Document clustering algorithms generally describe data using either Vector Space Models(VSMs) [277] or language models. The standard VSM model, on the other hand, suffers from excessive dimensionality, data sparsity, and semantic and morphological connection issues [278]. In the traditional VSM model, words are treated as atomic units, assuming that they are independent and that no idea of word similarity exists. As a result, the vocabulary in the textual data is enormous, and the vector is likewise high-dimensional. When we compute the similarity between documents, the volume of calculation is vast, and the similarity matrix is sparse in some cases. The VSM is likewise incapable of preserving word semantic and morphological associations. Despite the fact that

this has various advantages, including simplicity, robustness, and the fact that complex systems trained on small datasets outperformed simple systems trained on vast datasets, this basic approach just conducts word frequency statistics in the document. Several studies have considered the usage of word embeddings, such as word2vec, fastText, doc2vec, and others, for text representation in order to overcome the aforementioned issues.

This chapter employs distributed representations of words and phrases for vectors, allowing faster learning and recognition of more regular word representations. The Skip-gram or Continuous Bag-of-Words (CBOW) model, developed by T. Mikolov, K. Chen, G. Corrado, et al., is used to investigate language representation [279]. Using a basic neural network architecture, the CBOW and Skip-gram models can learn the text representation. The CBOW model predicts a word's nearby phrases based on its context. On the other hand, the Skip-gram model predicts the context of a single word. Because of their simplicity, these models can be trained on vast amounts of textual data.

T. Mikolov, K. Chen, G. Corrado, et al. demonstrated that the distributed representation of words well captures semantic and syntactic regularities [279]. Such representations possess a linear structure that integrates words meaningfully by adding their vector representations element-wise [280]. For instance, V("Paris")-V("France") +V("Spain") leads to a vector that is close to the V("Madrid"). Hence, these representations can preciously measure the words' similarities and resolve the synonym and semantic and morphological problems in the text. This chapter employs the Skip-Gram with Negative Sampling (SGNS) model for distribution word representation and improves text clustering accuracy.

Recently, low-dimensional representations have shown potential in overcoming the sparsity problem in online text clustering. Given the sparsity issue in short text grouping, deep neural networks provide significant gains in feature extraction. The deep clustering algorithm embeds data from high-dimensional to low-dimensional before performing the clustering method based on deep neural network feature extraction. Deep Embedded Clustering (DEC) is a deep clustering algorithm that uses an autoencoder to convert data to a lower-dimensional feature space[281]. It makes use of an allocation layer to increase the amount of clustering space available.

In order to enrich the document representation by retaining the semantic and morphological associations, this chapter introduced a word cluster-based modified tf-idf (WC_MTI) model, in which semantically associated word embeddings from the word2vec are supplemented with morphological information using KPCA. The vectorization of each text document was then based on the tf-idf values of the words it contained and the clusters in which they were detected. In addition, to address high dimensionality and sparsity issues and improve the clustering, we use a self-training technique that learns discriminative features using the WC_MTI model and autoencoder (AE) and then updates the encoder network weights using assignments from a clustering algorithm as supervision. The proposed model organizes documents

into topically compatible clusters by maintaining the semantic and morphological similarity between terms using the SGNS and low-dimensional vector representations.

The main advantage of the double-clustering strategy is that it reduces the feature space, which reduces the inevitable noise in the co-occurrence matrix of the original text terms. This reduced matrix is more robust and denser based on the feature clusters, providing a clearer representation of the underlying structure of the provided text.

5.2. Introduction to Related Methodologies

5.2.1. Kernel Principal Component Analysis (KPCA)

KPCA is a robust approach for feature extraction from possibly high-dimensional non-linear datasets [282]. The KPCA uses the Kernel Trick to map the original non-linear data into high-dimensional feature spaces, and PCA is conducted using iterative methods to estimate the sparse principal components in high-dimensional space. Embedding words based on morphological similarity may be viewed as a clustering challenge in a high-dimensional space that can be solved using Kernel PCA.

Given a vocabulary V of words w_i, a string similarity measure S and a non-linear kernel function K, a $|V| \times |V|$ word similarity matrix M is generated.

$$M_{ij} = K(S(w_i, w_j)) \tag{5.1}$$

Because column vector k_i of M may be seen as a $|V|$-dimensional representation of w_i, centering this kernel matrix [283] produces a feature space representation of words in V. PCA in this feature space then enables for the selection of the first $d < |V|$ non-linear principal components v_i, which allows for the projection of word vectors into lower-dimensional spaces. Using projection matrix

$$P = \left[\frac{v_1}{\lambda_1}, \ldots, \frac{v_d}{\lambda_d}\right] \tag{5.2}$$

formed by picking d eigenvectors v_1 to v_d corresponding to the largest eigenvalues, λ_1 to λ_d, the d-dimensional projections are

$$e_i = P^T k_i \tag{5.3}$$

5.2.2. Word2vec

A word2vec is a two-layer NN to map each token in a dataset to a low-dimensional vector representation in a fixed-size vector space. Two model architectures of word2vec can be used for computing continuous word vector representations: (1) CBOW and (2) Skip-Gram model. Both models have been proposed to

learn word embeddings while capturing semantic and synaptic information among words [279]. In both architectures, there is a target word and a set of context words. While the CBOW attempts to predict the context word based on its nearby words, the Skip-gram estimates the surrounding words of a target word [284]. T. Mikolov, K. Chen, G. Corrado, et al. proposed Skip-gram model-based distributed representation for words [280]. The primary benefit is that the representation of related terms in the vector space is very close. This method was used in several studies that involved the use of statistical language models, such as [279], as well as several other Natural Language Processing tasks such as word representation analysis [279,285], parsing [286], named entity recognition, labeling [287], etc.

Compared to the CBOW, the Skip-gram model can manage a large amount of data and identify infrequent terms conveniently in textual data processing. This chapter used the SGNS for distributed word representation [288]. Negative sampling facilitates modifying a small percentage of the weights for each training set instead of all of them. The skip-gram provides the probability of nearby words when given the word. The probability of predicting a word w_i is as follows provided the word w_j:

$$p(w_i|w_j) = \frac{exp(v'^{T}_{w_i} v_{w_j})}{\sum_{w=1}^{V} exp(v'^{T}_{w} v_{w_j})} \tag{5.4}$$

where v' and v are the target and context vector representation of the word w_i and the vocabulary size is V.

Skip-gram takes as input a one-hot encoded vector of size V based on the input word to estimate the context words within a window of fixed size C, as presented in Figure 5.1. For a set of words w_1,…, w_T, the purpose of the Skip-gram model is to maximize the following predictive accuracy:

$$I_{SG} = \frac{1}{V} \sum_{i=1}^{V} \sum_{-C \le t \le C, t \ne 0} log\ p(w_{i+t}|w_i) \tag{5.5}$$

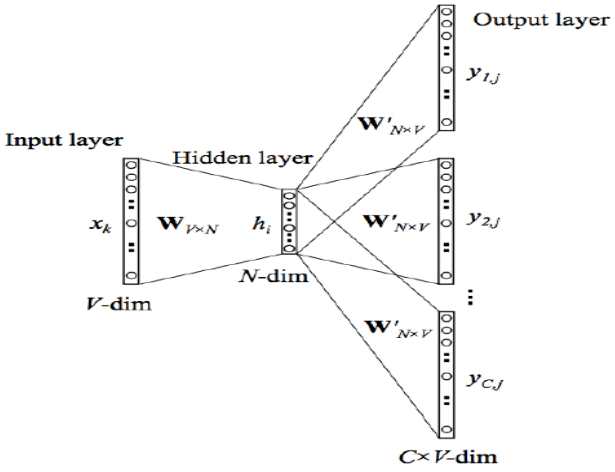


Figure 5.1 The Skip-Gram Model [289]

The formulation is costly since the complexity of calculating $log\, p\,(w_i|w_j)$ is dependent on the total terms in the vocabulary V. The use of a Negative-Sampling [290] is an efficient solution, which significantly reduces the computational complexity of $log\, p\,(w_i|w_j)$. This approximates the equation using sigmoid functions and k randomly sampled words, called negative samples. The resulting objective is given as:

$$I_{SGNS} = \frac{1}{V}\sum_{i=1}^{V}\sum_{-C\le t\le C, t\neq 0} log\ \sigma\left(v'^{T}_{w_i} v_{w_j}\right) + k\mathbb{E}_{t\sim P(w)}[log\ \sigma\left(-v'^{T}_{t} v_{w_j}\right)] \qquad (5.6)$$

5.2.3. Vector Space Model (VSM)

One of the easiest ways to represent a document is in terms of the words it contains. The sequencing of the words is ignored in such representations of texts, and just the occurrence remains. In this chapter, the Vector Space Model is used for such document representation. It [277] is a statistical model representing text information for several NLP tasks. Such models contain information about each document's essential terms and how many documents contain a particular term. Statistically, as shown in Figure 5.2 in the vector space model, each document (D_1,D_2, D_3,... D_n) is represented by a vector of terms ($\overrightarrow{D1}$, $\overrightarrow{D2}$, $\overrightarrow{D3}$.....$\overrightarrow{Dn}$) extracted from the document, having corresponding weights w_{ij}, describing the importance of the terms W_i in the document D_j. Each term represents a feature, and there are many ways to determine the weight of the terms. A common approach for weight calculation is to use the tf-idf method. The tf-idf

is a product between term frequency (tf) and inverse document frequency (idf) [291]. The tf determines how often in a document d a word t appears. IDF accesses the significance of a word t in the sets of documents D. For each document, commonly occurring terms such as 'a', 'an',' the' can have high tf but small idf as it does not contribute significant value to an existing document. However, given that each term in the corpus constitutes a feature, such representations suffer from sparsity and high dimensionality, which may not contribute to useful clustering [292].

$$tf - idf\ (d,t,D) = \ tf(t,d) * idf(t,D) \tag{5.7}$$

$$tf(t,d) = \frac{number\ of\ occurrence\ of\ t \in d}{the\ total\ length\ of\ d} \tag{5.8}$$

$$idf(t,D) = log\frac{number\ of\ documents \in D}{number\ of\ documents\ of\ Dconatining\ term\ t} \tag{5.9}$$

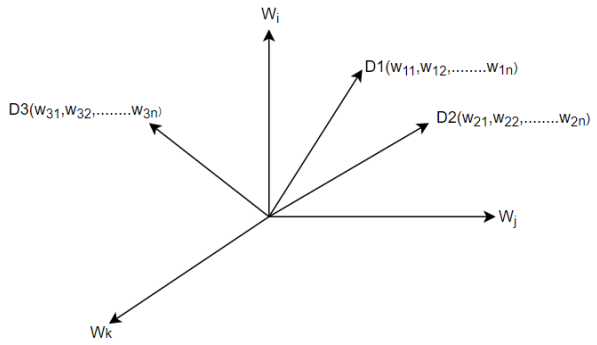


Figure 5.2 Schematic diagram of Vector Space Model

5.2.4. k-means Clustering

After obtaining the text representation in the vector form for the collected data, the next step is to group similar words to identify the topics in the data set. The k-means [293] is a well-known and often used unsupervised clustering technique. It distributes the $'n'$ data points into $'k'$ clusters to minimize the sum of distances of each data point within the cluster and to their cluster's center. Following the selection of k initial centroids, a cluster is allocated to each document based on a distance metric, after which k centroids are recomputed. The process is repeated until an optimum set of k clusters is produced based on an objective function. So far, because of its efficiency, k-means is amongst the most common text clustering algorithms [294,295]. This algorithm aims to search for k clusters based on the following objective function optimization, given a data set $X = \{d_1, \dots, d_n\}$ containing n documents:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} sim(d_i, c_j) \quad (5.10)$$

Depending on the $sim(d_i, c_j)$ definition, J is either minimized or maximized. The centroid of the cluster C_j is defined by c_j. The similarity between document d_i and centroid c_j is evaluated by $sim(d_i, c_j)$. The cosine similarity is one of the most widely implemented measures [296] for calculating the similarity between two documents d_i and d_j:

$$cos\,\theta = \frac{d_i d_j}{||d_i||\,||d_j||} \quad (5.11)$$

In our case, the objective function is maximized. This chapter implemented cosine distance with k-means as it is faster, deals better with sparse matrices, and calculates distances independent of texts' lengths. Thus, a short text with fewer terms can be considered very similar to a longer text with several terms. In addition, a document with a large number of words has a very high dimensional space. Therefore, we compute the distance using the cosine metric rather than the Euclidean distance.

5.2.5. Deep Embedded Clustering

While deep learning is a relatively new discipline of machine learning, it only became popular in the first decade of the twenty-first century. The main method of deep learning is to utilize the distance value as an input signal to fine-tune the weight values in order to reduce the loss values associated with the current instance. The optimizer does this modification by implementing the Back Propagation method, the essential deep learning algorithm.

Deep clustering automates the process of extracting features, reducing dimensionality, and clustering. J.Xie, R. Girshick, A. Farhadi has described a deep embedded clustering method [281]. The Stacked Autoencoder (SAE) model is the foundation for the DEC model. Two stages of the DEC technique are parameter initialization using deep auto-encoder and parameter (clustering) optimization. DEC is a parameterized non-linear representation from the original data space X to a lower-dimensional space Z, with the clustering aim optimized.

An autoencoder is a data compression method that learns data-dependent, lossy compression and decompression properties from the sample. When autoencoders are defined, the compression and decompression functions are often accomplished using neural networks. An autoencoder NN is an unsupervised learning method that uses backpropagation on a series of unlabeled training samples and defines the output value to be the same as the inputs [297]. The basic framework of AE is a feed-forward NN of two layers having one hidden layer. It includes the encoding and decoding processes. After feeding the feature vector x into the encoder, non-linear modifications are performed, the activation function is processed, and a coding result is obtained. They compress the input data using an encoder function and

recreate it using a decoder function. Autoencoder attempts to limit the divergence of output z from the input x. The process of an autoencoder with a single hidden layer is explained as follows:

$$y = f(Wx + b) \tag{5.12}$$

$$z = f(W'y + b') \tag{5.13}$$

Where y and z represents encoding and decoding results, respectively

f is a non-linear activation function,

W, b, and W', b' are encoding and decoding weight and bias matrix

We may utilize an unsupervised layer-wise training approach to build deep-SAEs that are then fine-tuned to reduce reconstruction loss to fully use deep learning's remarkable capacity to learn a latent input representation [298]. The SAE structure is depicted in Figure 5.3.

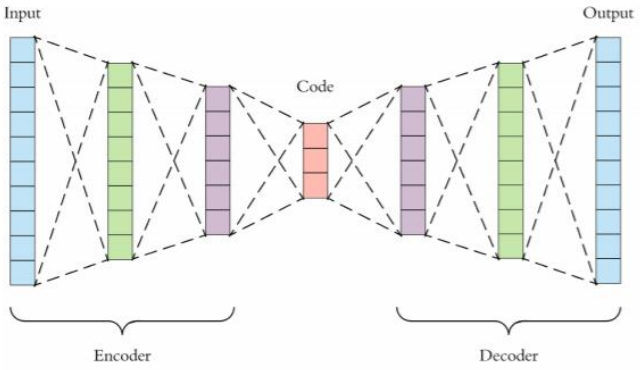


Figure 5.3 Structure of SAE Model [299]

5.3. Proposed Methodology, WC_MTI

Figure 5.4 depicts the logical workflow of the proposed approach. The algorithm for the WC_MTI is presented in Algorithm 5.1.

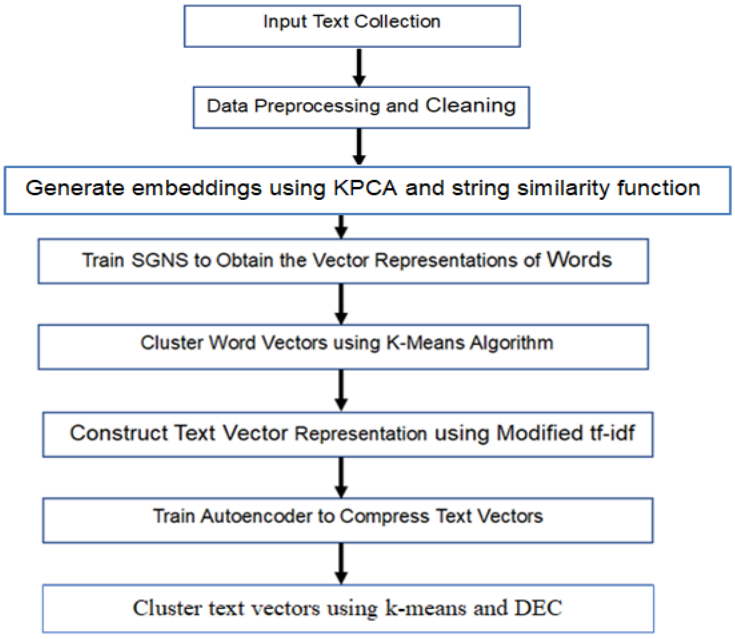


Figure 5.4 Logical Workflow of the Proposed Approach

Algorithm 5.1: WC_MTI

Input: the corpus of texts, k and k' be the number of word clusters,

Output: Clustering results

calculate vector representation of words by KPCA-based Skip-gram with Negative sampling

the k partitions of word clusters are obtained by k-means using cosine similarity

for each text, D_i do

for each cluster, C_j do

for each token in C_j do

Calculate the weight w_{ji} of C_j with regards to D_i using modified tf-idf

End for

End for

Compute text vector representation

End for

Compress the vector representation using deep autoencoders

Conduct the self-training and k-means method to output the final k' clusters

5.3.1. Datasets Collection

Tweets are collected using a Twitter API to extract the trending topics related to mental health, containing specific keywords and hashtags. Tweets were collected in the form of a CSV file. Besides the text message and the presence of particular hashtags, the obtained records include other attributes, such as date, information about retweets, geographic location, tweet id, link to the tweet, etc. In our research, we only require hashtags and text data. Therefore, we did not use any other information. Since mental health topics are linked to many different activities, the tweets collected are suitable for comparing document clustering. We recorded tweets between September 16, 2020, and December 15, 2020, containing the hashtag #mentalhealth. Thereby, a total of 2,92,432 tweets were collected.

Reddit comments were taken from [300]. In order to characterize trends in 15 support groups for mental health (e.g., r/schizophrenia, r/SuicideWatch, r/Depression), the authors developed the Reddit Mental Health Dataset. The Reddit forum is commonly used, grouped by the subreddit list, to discuss specific topics or themes, so it is suitable for this analysis. From 2018 to 2020, the entire dataset includes posts from 8,26,961 individual users. Since it is publicly available and discusses the health theme, we chose this data collection. Reddit data is particularly relevant to the research since its posts have a larger range of character lengths than Twitter data.

5.3.2. Data Pre-processing and Cleaning

A series of pre-processing steps must be followed before using textual data for information extraction purposes. The online captured data transformation into a clean and easy-to-analyze format requires a few steps. Initially, we started preparing the data by considering the textual information of the tweets and Reddit comments as the basic unit. Next, we perform case conversion and standardize the characters for both datasets. We remove the stop-words using Natural Language Toolkit (NLTK), with some additional elements such as "RT". We use the well-known Snowball's stemming algorithm for stemming, and we only considered terms with more than two characters.

Removing URLs: Most of the URLs in the collected tweets point to blog posts shared by users; therefore, we removed the URLs.

Lowercasing alphabets: The collected data possibly consists of several terms whose spellings are the same but letters in different cases. For instance, the collected tweets contain the terms' Pollution' and 'pollution' in different letter cases that would be considered different while clustering. Such cases may affect the clustering results, and hence every single letter is changed to lowercase.

Eliminating stop words: Stop words are very frequent words in textual data and carry almost no semantics. A list of stop words includes articles, prepositions, conjunctions, etc. As stop-words do not contribute any use in our research, we removed stop-words using NLTK. We eliminated the noise like various operators, symbols like @, ?, $, %, etc., and punctuation marks. We also pulled the non-ASCII characters, numeric digits, and unnecessary white spaces.

Stemming: Another step is stemming. It is a procedure that connects words of similar semantic with the same roots. For example, 'help' and 'helped' have the same root form, 'help'. The stemming process reduces the suffixes, prefixes, plurals, and verbal variants to a single form. Our work used NLTK's Snowball Stemmer to reduce all stemmed words to their root form.

5.3.3. Data Preparation

As Twitter data does not contain any Ground Truth label, we provided the topic label to the top 50 high-count hashtags. We used these topic labels as the Ground Truth labels. Since we have collected data using #mentalhealth, most tweets have hashtags related to health. Therefore, we removed the hashtags #mentalhealth and #health from the data set. Next, the top 50 hashtags, with at least 200 related tweets under one label, were filtered. We considered the related hashtags under one label. For example, "#coronavirus" is one of the top hashtags, and therefore we have included other hashtags such as "#coronavirus", "#covid19", "#covid", etc., under the same label. In case a tweet contains multiple hashtags, we assigned it to the more appropriate truth label. We summarize the results of the hashtags with assigned labels and their counts in Table 5.1.

Table 5.1 Hashtags from the Dataset with Labels

Label#	Hashtags	Label#	Hashtags
0	#covid, #coronavirus, #covid19, #pandemic	5	#vaccines, #vaccinessaveslives
1	#young, #people, #child	6	# love, #mindfulness, #selfcare, #life, #motivation, #well, #therapy, #recovery, #help
2	#lockdown, #school, #work	7	#suicideprevention, #suicideawareness
3	#mentalhealthawareness, #mentalillness, #mentalhealthmatters, #psychology, #issues	8	#socialmedia

4	# anxiety, #depressed, #sad, #sadness, #ptsd, #stress, #addiction, #bpd, #bipolar	9	#spiritualguruofmillions, #jesus, #prayer, #jesusheals

Reddit data on 15 unique mental health support groups [300] were analyzed. We removed punctuations and standard stop words to tokenize the Reddit data. Most posts have the same word, so we removed the subreddit page titles from the corresponding subreddit page's data. We did not need to change the text for the Reddit data because we used the subreddit pages as Ground Truth labels.

5.3.4. Distributed Word Representation

After the pre-processing step, each document is now a series of text elements. When working with the text representation methods, we need to represent text in a mathematical format to solve problems. In other words, the textual data need to be converted into an integer representation. In this chapter, the SGNS model was initialized with the pre-trained KPCA embeddings. Now, each word will be represented by a continuous d-dimension vector instead of a discrete and sparse vector, and its relationship to other terms will capture their meaning. The explanation for this is that if two selected words have the same context in training time, conceptually, the weight of the network would be close to each other for these two target words and hence their corresponding vectors. Thus, we get a distributional representation for each term in the corpus. There are various attributes or hyper-parameters of this model.

We experimented with size or dimension hyper-parameter over a range of values, 50-300, and found the accuracy to be maximum at 100 dimensions for both Twitter and Reddit datasets. The general view of the result for the Twitter dataset is shown in Figure 5.5, which exhibits that the maximum peak is at 100 dimensions, followed by 50 dimensions. Next, we fixed the value of 5 for window size as this size could capture all the tokens of most documents. We found that the chosen hyper-parameters were optimal in performance using the three performance measures. For the rest of the hyper-parameters, the default values of the word2vec model provided by gensim 3.8.1 python package are used. This representation will eventually lead to the identification of semantic and morphologically coherent documents.

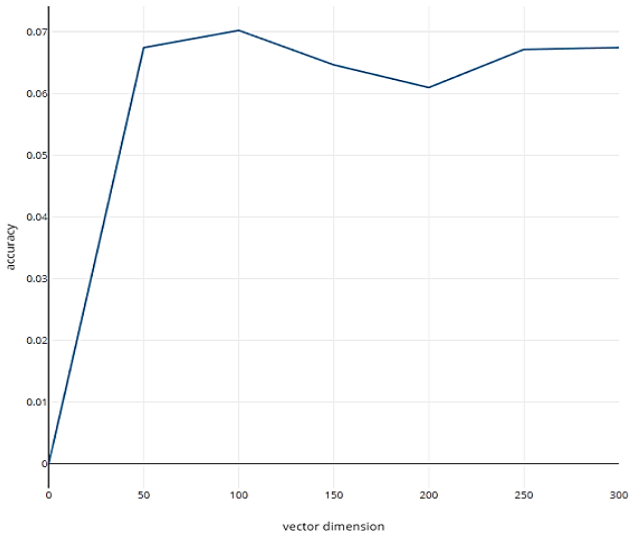


Figure 5.5 Word2vec Dimension-Accuracy Graph

5.3.5. Clustering Word Embeddings

Next, we implement clustering over the word vectors to classify various groups of words that match together semantically and morphologically using the k-means algorithm. As a result, we obtain the predefined number of word clusters. The semantic and morphological cluster of terms can be obtained by clustering word embeddings. This chapter implemented cosine distance as a distance metric as it is faster, deals better with sparse matrices, and calculates distances independent of texts' lengths. Thus, a short text with fewer terms can be considered similar to a longer text with several terms. In addition, a document with a large number of words has a very high dimensional space. Therefore, we compute the distance using the cosine metric rather than the Euclidean distance. We used the cosine metric from NLTK in k-means implementation.

5.3.6. Feature Vectors Construction

Our work presents a new text representation method that uses the word clusters obtained in the preceding step. In our method, the new features are word clusters, so the dimension of new vectors is the same as the number of word clusters. In this manner, the proposed method reduces vector representation's dimensionality, resulting in a modified tf-idf based on feature clusters. Moreover, semantic and morphologically connected words are grouped by each cluster so that regardless of the word present in a particular document T but is semantic and morphologically related to words that exist in document T, the weight of that word is included in the total weight of the respective cluster in the document. The tf-idf weighting is altered as follows:

$$w_{ij} = \frac{tf_{ab} * idf_a}{\sum_{w \in T_b} tf_{ab} * idf_a} * AF \quad (5.14)$$

The equation defines the weight allocated to the word cluster 'C_a ' in the document 'T_b'. tf_{ab} is the frequency of C_a in T_b, which is the total tf weights of the terms w present in cluster C_a and document T_b as referred to the equation(5.15), idf_a is the idf weighting of C_a, i.e., the total idf weights of words w belonging to C_a mentioned in equation (5.16). The denominator is used for vector normalization.

AF is the adjusting factor that ensures that both lengthy and short documents are given equal consideration. It enhances the importance of term frequency in a text and penalizes terms that appear in a document less frequently with a larger term frequency weighting. AF can be calculated, as mentioned in equation (5.17), which is the aggregation of the term frequencies of the unique terms.

$$tf_{ab} = \sum_{l=1}^{n} tf(w_l \in C_{a,} T_b), a \in \{1, \ldots \ldots, k\}, b \in \{1, \ldots, m\} \quad (5.15)$$

$$idf_a = \sum_{l=1}^{n} idf(w_l \in C_{a,}), a \in \{1, \ldots \ldots, k\} \quad (5.16)$$

$$AF = ln(L|1 + L - tf(t, d)) \quad (5.17)$$

The updated document vector representations add various enhancements to the standard VSM. In the text modeling, the tf-idf method weights the words representing the influence of a word on the document and produces a score, whereas word embedding models produce mapping for each word by preserving its semantic and morphological links. Our approach incorporates the benefits of both methods and increases clustering effectiveness. It not only overcomes the problem of dimensionality but also holds semantic and morphological by using the word clusters. Since we trained on the words and obtained the word clusters, we employed the word clusters for document representation. The main improvements are that we replaced the words with the word clusters as new features. However, there is no major difference in form, but there is a change in the training method and improved clustering efficiency. Moreover, based on the new weighting, the yielded representation is significantly less sparse than the conventional representation.

5.3.7. Compressing Feature Vectors

Next, a deep autoencoder was trained over the text vectors to compress their dimensionality. The layers were lined to construct the encoder and the decoder component of the autoencoder. For tweet vectors, the output size of the encoder layers was successively reduced from 165 to 12, and the output value of the decoder layer was successively increased from 12 to 165. Whereas for Reddit comments vectors, the dimensions were reduced from 300 to 30 by the output value of the encoder and decreased from 300 to 30 as the output size of the decoder. According to [301], the purpose of the increased number of layers in the model was to see if a deep learning model performs better with larger networks. The stacked encoder

model was trained with a batch size of 128 using the Google Colab pro version. Except for the additional dimensions in the encoded layers of the input data space, the hyperparameters of the neural network model were trained using default values. With tuned hyperparameters, the network might deliver better outcomes. The model may, for example, have been trained with various batch sizes. However, with such a huge network, training the encoder model in Google Colab pro takes a very long time.

The activation function in all layers was Rectified Linear Unit (ReLU), except for the end layer, which was a linear activation function. Adam was chosen as the optimizer for autoencoder training because of its fast convergence capabilities, and the loss function was Mean Squared Error. The decoder layers were deleted after the autoencoder was trained, and the document vectors were introduced into the model to generate the respected compressed representation. Although having less sparsity, the compressed vectors are an alternative representation of text vectors. This is supported by our model reconstructing the initial document vector accurately from its compressed forms with a slight loss of 0.056 for tweet vectors and 0.088 for Reddit vectors, using binary-cross entropy as the loss function. This significantly decreases the data's dimensionality. With larger corpora, several clusters of more than 300 words could be generated. The document vectors will also have very high dimensionality in such scenarios. DEC is implemented in the same way as J. Xie, R. Girshick, A.Farhadi [281].

5.3.8. Clustering with Kullback–Leibler (KL) Divergence

Soft labeling is one of the most important aspects of the DEC model. Soft labeling is the process of assigning an estimated class to each of the data samples that may be refined repeatedly. The low-dimensional representation of the non-linear mapping feature embedded from modified tf-idf was obtained after pre-training with an autoencoder, and then a clustering algorithm was applied to the obtained features.

Inspired by previous work on self-training mechanisms, the clustering layer is defined based on clustering loss [281]. The technique can use the high confidence clustering assignments as soft labels to assist the optimization process. The clustering performance may be improved iteratively in this way. We first initialize clustering centers and iterate the following steps:

(i) compute the probability for each clustering center point;

(ii) compute the auxiliary target distribution and deploy it as an encoder's objective network. As illustrated in Figure 5.6, the clustering centers and network weights are repeatedly revised until the convergence condition is fulfilled.

For step (i), as motivated by the prior t-SNE method [302], the similarity between the center point c_j and its surrounding points c_j follows the distribution. as a kernel, we utilize the Student's t-distribution as a

kernel measure to assess the similarity between the center of the cluster c_j, and the node embedding representation z_i, quantified by q_{ij}:

$$q_{ij} = \frac{(1+||z_i - c_j||^2/\alpha)^{-1}}{\sum_k (1+||z_i - c_k||^2/\alpha)^{-1}} \tag{5.18}$$

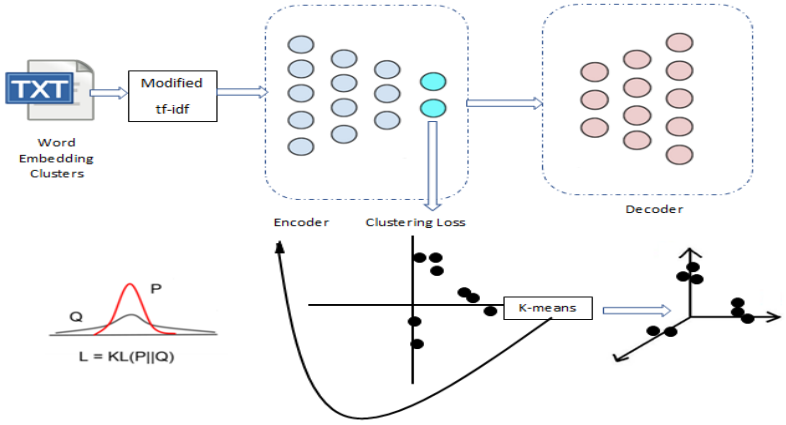


Figure 5.6 Self-training Mechanism

The auxiliary target distribution has higher accuracy than q_{ij}. In step (ii), similar to [281], auxiliary target distribution P is used, which is computed by normalizing by frequency per cluster as follows:

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_k q_{ik}^2 / \sum_i q_{ik}} \tag{5.19}$$

As can be seen, the target distribution P raised Q to its second power to give high-confidence data samples more importance in the distribution. By decreasing the distance between auxiliary target distributions P and soft assignment, the clustering layer refines centroids by learning from highly confident assignments. The soft assignment distribution q_{ij} is expected to be consistent with the auxiliary target distributions P. To evaluate divergence between the two distributions P and Q, we use the KL divergence as a loss function defined as follows:

$$L = KL(P||Q) = \sum_i \sum_j p_{i,j} \, log \frac{p_{ij}}{q_{ij}} \tag{5.20}$$

DEC optimization is challenging. The objective is to learn the original vector representation and cluster assignment simultaneously. Here an iterative refining cluster technique with an auxiliary target distribution generated from the present soft cluster assignment is used. The soft assignment is matched to

the target distribution to train the model. This model uses a k-means model and a pre-trained auto-encoder to create a new model that accepts pre-processed textual data as input and outputs both predicted clustering classes and decoded input data records.

5.4. Experiments and Results

Several experiments were carried out to examine the feasibility of the proposed model. First, we compared the performance of the proposed text representation technique, WC_MTI, with the existing approaches with a different number of epochs. Next, the experiments were carried out to evaluate the impact of vector representations by different text representation methods on the clustering results.

5.4.1. Comparison Methods

The detail of the comparison methods are as follows:

- Tf-idf: traditional tf-idf matrix is used.
- Average word2vec (TR1): The average word2vec was obtained by computing for every term in each document the element-wise average of the word vectors, resulting in a compact feature vector.
- tf-idf weighted word2vec (TR2): For each term in a document, TR2 was calculated by finding the tf-idf weighted average of the word vectors.

5.4.2. Evaluation Measures

We have used external measures such as Mutual Information (MI) and Rand Index (RI) for evaluating document clustering methods. The basis of the external metrics is the output of the clustering system and ground truth labels [303]. Such measures do not rely on the absolute values of the labels. M. P. Naik, H. B. Prajapati, and V. K. Dabhi [304] compared the internal methods such as F1, recall, and precision with the methods that use different feature sets. Also, there is a need for a fixed bound for evaluation by external clustering measures. Therefore, we have chosen Normalised Mutual Information (NMI) and RI, which generate values between 0 and 1.

Mutual information computes the mutual dependence of two or more random variables. It measures the degradation of one discrete random variable's uncertainty and provides information about another discrete variable [305]. In terms of cluster distribution, MI is unbiased and symmetric [303]. High mutual information indicates a better performance. The MI for two discrete variables X and Y, is given by:

$$MI(X,Y) = \sum_{\Omega X} \sum_{\Omega Y} p(x,y)\ log_2 \frac{p(x,y)}{p(x)p(y)}\ d_x\ d_y \tag{5.21}$$

where:

ΩX and ΩY represents sample spaces of X and Y;

$p(x, y)$ represents the joint probability mass functions of X and Y;

$p(x)$ and $p(y)$ signify the marginal probability mass functions of X and Y.

5.4.2.1. Normalized Mutual Information (NMI)

NMI is a widely used measure that normalizes the mutual information between two random variables to have normalized values between 0 and 1, with 1 indicating mutual information and 0 being disagreement. The formulation for NMI is described as follows:

$$NMI(X,Y) = \frac{MI(X,Y)}{\sqrt{H(X)+H(Y)}} \tag{5.22}$$

$$H(X) = -\sum_{i=1}^{n} p(x_i)\, log(p(x_i)) \tag{5.23}$$

where:

X denotes class labels;

Y denotes cluster labels;

$H(X)$ and $H(Y)$ are the entropies of X and Y, respectively.

5.4.2.2. Rand Index (RI)

RI is one of the most widely accepted cluster validation measures [306]. It computes how similar the clusters are to benchmark labels. It generates values between 0 and 1, with 1 being identical labels and 0 denoting random labeling. For a set of elements, $S = \{s1, s2 \ldots, sn\}$ with two partitions, $X = \{X1, \ldots, Xr\}$ and $Y = \{Y1, \ldots, Ys\}$, the RI represents the proportions of agreement between X and Y. Rand Index, RI, can be calculated as:

$$RI(X,Y) = \frac{a+b}{a+b+c+d} \tag{5.24}$$

where:

a represents total pairs in S that are correct similar pairs in X and Y;

b represents total pairs in S that are correct dissimilar in X and Y;

$a + b$ gives the total agreements between X and Y;

c defines the total pairs in set S such that they belong to the same class of X but to distinct classes of Y;

d represents the total pairs in set S such that they belong to different classes of X and the same class of Y;

c and d together give the total disagreements between X and Y.

It has been shown that external measures comparing two independent clusterings should have constant baseline property, implying that they can compare clustering methods with different clusters. Measures without this property tend to select clustering with more clusters [307]. L. Hubert and P. Arabie[308] pointed out the same problem for the RI and proposed the Adjusted Rand Index (ARI) as follows:

$$ARI = \frac{RI - E\{RI\}}{max\{RI\} - E\{RI\}} \tag{5.25}$$

where $E\{RI\}$ represents the expected value of RI, which generates the values over the range [0-1], with 0 representing unidentical partitions and 1 being identical. It is often required to measure the agreement between the clustering partition and that of external criteria.

5.4.2.3. Accuracy

We also used clustering accuracy (ACC) as the evaluation index. Equation (5.26) shows the definition of clustering accuracy:

$$ACC = \frac{\sum_{i=1}^{N} \delta(l_i = map(c_i))}{N} \tag{5.26}$$

Where:

$\delta()$ is an indicator function;

c_i is the cluster label of x_i;

map () convert the cluster label c_i to its group label using the Hungarian algorithm;

y_i is the true group label of x_i.

Experiments are performed to validate the proposed approach's effectiveness relative to the conventional VSM, average word2vec, and tf-idf weighted word2vec.

5.4.3. Number of Epochs

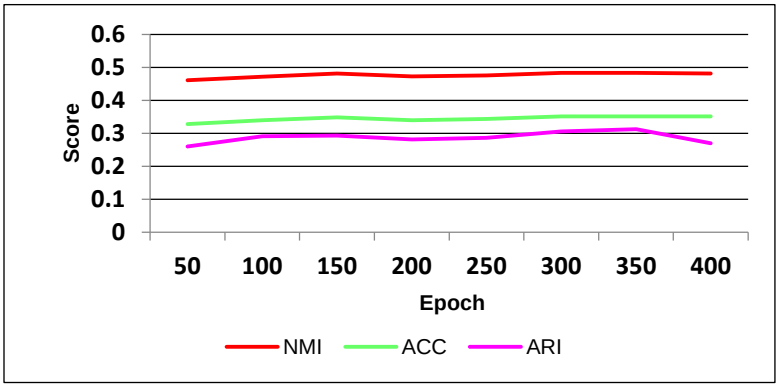
While applying a neural network, the number of epochs or iterations hyper-parameter plays a critical role in the model's performance. Suppose the number of epochs is less than the optimum value; in that case, the model will not be trained thoroughly, and too many epochs will lead to overfitting.

Firstly, we analyzed the performance change for all the text representation models with the number of epochs. Further, for clustering, we used the k-means algorithm to give reliable results for the text representations. We trained each text representation model with epoch values in the range of 50 and 400, incrementing 50 in every iteration and evaluated against the Ground Truth labels. We ran each text representation method 10 times and considered the average value for each evaluation measure. Table 5.2 summarizes the optimal number of epochs produced by each method. Figure 5.7(a)-(c) shows how ACC,

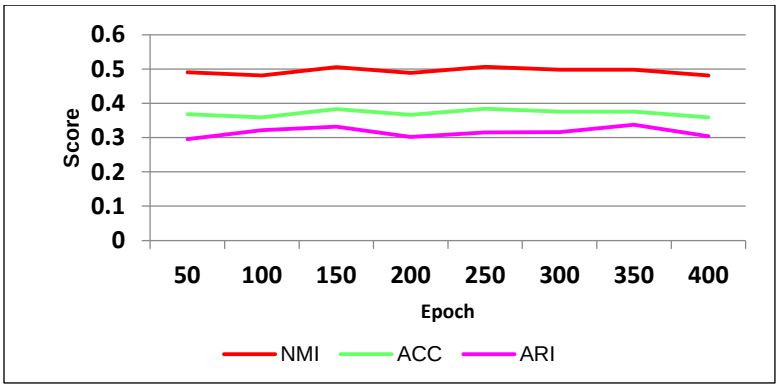
NMI, and ARI scores vary with the number of epochs on Twitter data. The analysis of Reddit data is shown in Figure 5.8(a)-(c).

Table 5.2 Optimal Number of Epochs for Each Text Representation Method

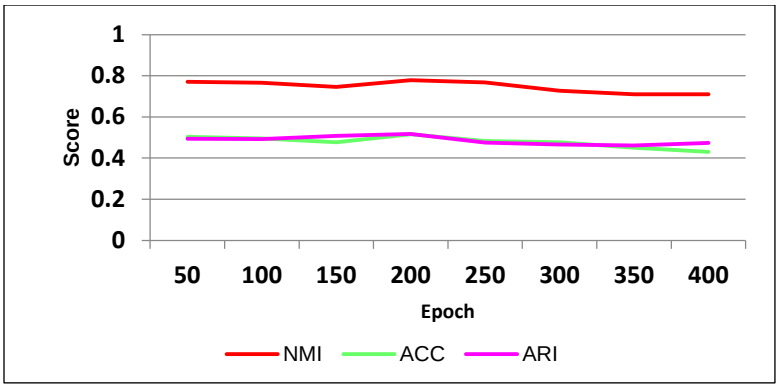
	Text representation Method	Optimal No. of Epochs
Twitter	TR1	350
	TR2	250
	WC_MTI	200
Reddit	TR1	100
	TR2	50
	WC_MTI	100



(a)

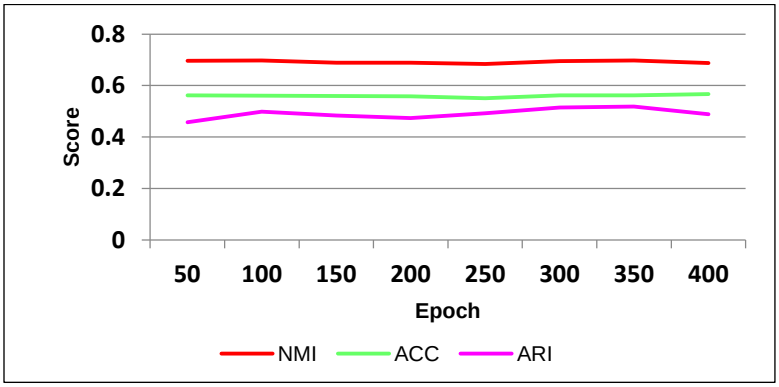


(b)

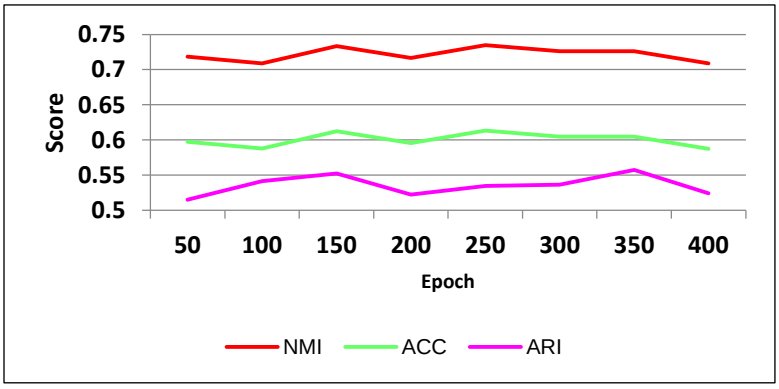


(c)

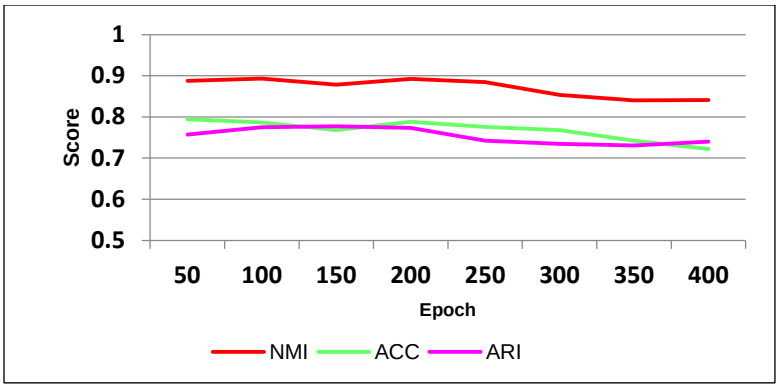
Figure 5.7 Plots of Three Evaluation Measures, i.e., NMI, ACC and ARI for Different Text Representation Methods (a) Average word2vec (b) tf-idf weighted word2vec and (c) WC_MTI on Twitter Data



(a)



(b)

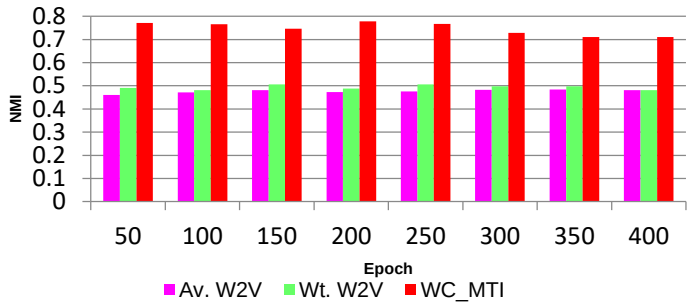


(c)

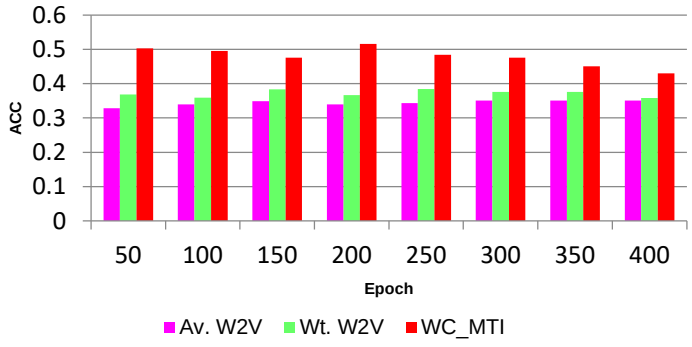
Figure 5.8 Plots of Three Evaluation Measures, i.e., NMI, ACC, and ARI for Different Text Representation Methods (a) Average word2vec (b) tf-idf weighted word2vec and (c) WC_MTI on Reddit Data

It is clear from Figure 5.7 that WC_MTI had peak performance at around 200 epochs for Twitter data. The average word2vec method achieved better NMI, ACC, and ARI values with more epochs, i.e., about 350. The word2vec technique weighted by tf-idf performed better than the average word2vec process. Turning to the Reddit dataset outcome, the best result was again provided by our proposed method. The TR2 method showed superior performance than the average word vector technique and required fewer training epochs. These findings appear close to the Twitter data set results. On the Reddit data, each text representation technique required fewer epochs to achieve optimal performance. It is clear from the observation that:

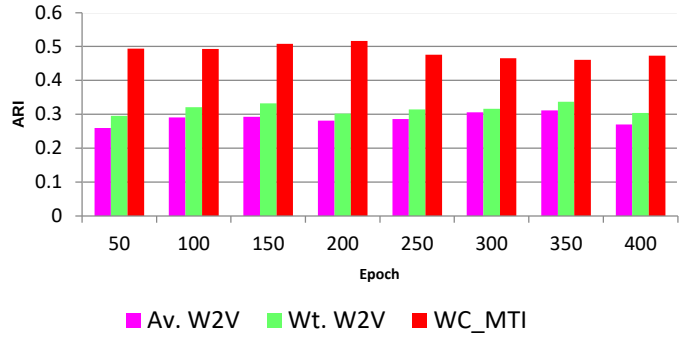
- Our method consistently provided better results than other approaches.
- However, our method required more training epochs than TR2 when the data set had larger documents, i.e., Reddit comments. In particular, weighting by tf-idf values boosted the performance and allowed fewer epochs for training.
- It is clear from Figure 5.7 and Figure 5.8 that the number of epochs for training the Twitter dataset was more than the Reddit dataset supporting the evident pattern that more training epochs are needed for shorter documents to achieve the optimal results.



(a)

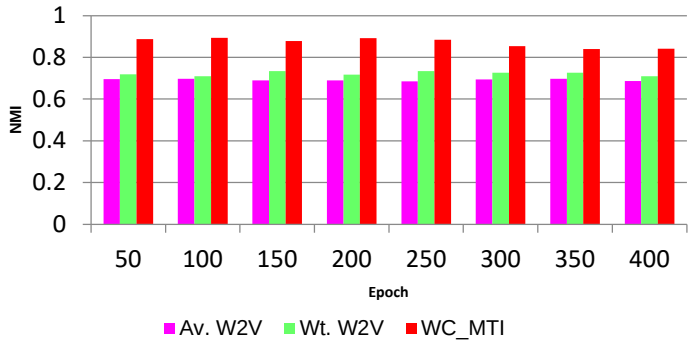


(b)

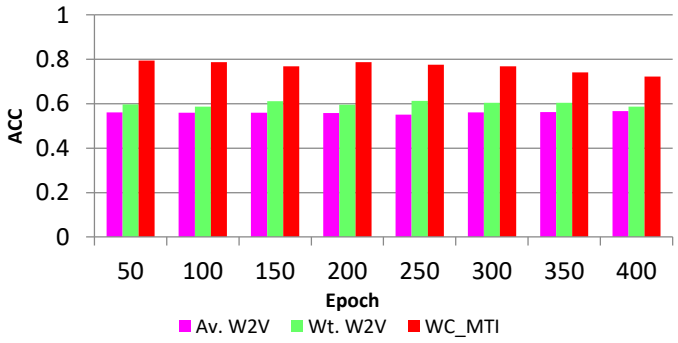


(c)

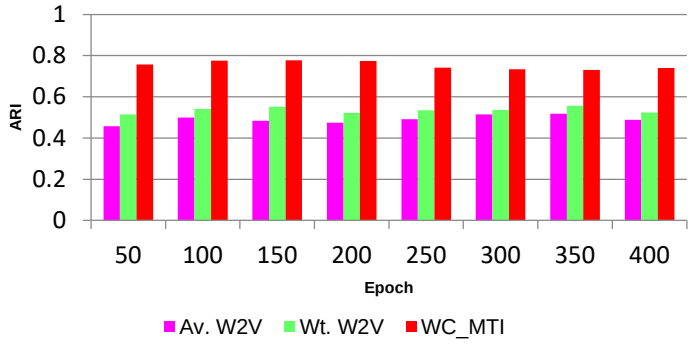
Figure 5.9 Plots of Performance of Text Representation Methods, i.e., Average word2vec (Av. W2V), tf-idf weighted word2vec (Wt. W2V), and WC_MTI on (a) NMI, (b) ACC and (c) ARI on Twitter Data



(a)



(b)



(c)

Figure 5.10 Plots of Performance of Text Representation Methods, i.e., Average word2vec (Av. W2V), tf-idf weighted word2vec (Wt. W2V), and WC_MTI on (a) NMI, (b) ACC and (c) ARI on Reddit Data

5.4.4. Performance Evaluation with Clustering Measures

The mean value of NMI, ACC, and ARI for each method is set out in Table 5.3 for both data sets. The number of epochs was fixed to be the same as mentioned in Table 5.2. Figure 5.9 and Figure 5.10 show the projections for the evaluation measures over Twitter and Reddit data, respectively. The introduction of

additional fine-tuning and self-training in low-dimensional representations in the proposed method leads to improved cluster quality. It is evident from these figures and Table 5.3 that our method was noticeably ahead of the remaining text embedding methods on all three measures.

Table 5.3 Performance of Different Text Representation Methods with DEC in Terms of ACC, NMI, and ARI

Dataset	Method	ACC	NMI	ARI
Twitter	**tf-idf**	0.424	0.469	0.108
	Average word2vec	0.344	0.476	0.287
	tf-idf weighted word2vec	0.371	0.494	0.315
	WC_MTI	**0.479**	**0.747**	**0.485**
Reddit	**tf-idf**	0.405	0.435	0.138
	Average word2vec	0.56	0.691	0.490
	tf-idf weighted word2vec	0.60	0.721	0.535
	WC_MTI	**0.768**	**0.871**	**0.753**

Lastly, we present the accuracy and NMI and ARI results of the model and comparison tests over the Reddit data. Figure 5.10(a)-(c) shows the NMI, ACC, and ARI values for various techniques on the Reddit data, indicating the proposed method outperformed the other text representation methods. Furthermore, the VSM weighted average word2vec approach provided better results than the average word2vec method. These results appear similar to the Twitter data set result. Compared to the average word2vec method, the weighted word2vec method constantly provided a performance boost.

We listed the results of all methods in Table 5.3. Compared to all comparison methods, WC_MTI produced the best clustering performance. More details are discussed as follows:

- Our proposed approach enhanced the average ACC by 89.62% and 39.24% on the Reddit and Twitter datasets.
- On the Reddit dataset, on average, NMI improved by 20.8% and 26.04% and ARI by 40.74% and 53.67%, compared to tf-idf weighted word2vec and average word2vec, respectively, while on the Twitter dataset, on average, the NMI was improved by 51.21% and 56.93% using WC_MTI and ARI by 53.96% and 68.98% compared to tf-idf weighted word2vec and average word2vec

respectively. The reason may be that our method employs less sparse vector representations and preserves semantic and morphological similarities.

- The tf-idf worked relatively well on Twitter data is much better than on the Reddit data suggesting that the Twitter data topics may heavily rely on keywords, considering that hashtags and user mentions are not alarming. Overall, the performance of Reddit data is far better than that of Twitter data, which is supported by the fact that it has annotated more discrete topics.
- Interestingly, the tf-idf method did not give comparable performance to the word2vec methods for both datasets, indicating the role of the semantic and morphological similarity of word2vec methods in clustering performance.
- Our method shows a drop in the evaluation measures after 200 epochs for both datasets while the other methods improved. One possible reason for this fall is that averaging document vectors for feature clusters may hold true over a threshold of words.
- Also, on the Twitter data set, there was a noticeable difference between the highest and lowest performance of the proposed method relative to the other methods. In contrast, a comparable performance was observed on the Reddit data set for all the methods, indicating that other methods handled the Reddit data set better than WC_MTI. Nevertheless, our method still performed best on all evaluation measures across both data sets.
- In addition, experimental findings indicate no substantial difference between the Tweets and the Reddit comments clustering performance, indicating the approach proposed here can be regarded as platform insensitive.

5.5. Chapter Summary

This chapter concerned document clustering that focuses on grouping topically homogeneous textual data. The existing text representation approaches required for clustering suffers from several drawbacks, such as high dimensionality and avoidance of semantic and morphological relations among words. The double clustering approach presented in the chapter aims to reduce dimensionality and sparsity of vector representations, preserving the semantic and morphological relations between words to improve cluster performance. The KPCA-based negative sampling of the skip-gram model, modified tf-idf method based on feature clusters, and deep autoencoders fine-tuned for clustering using self-training are the basis of our approach. Experimental findings also showed that the proposed method’s clustering performance was higher than the other existing methods, proving the proposed method's effectiveness. Since the current implementation of word2vec focuses on the uni-gram, the approach used in the chapter has the scope of improvement by learning embeddings for more than one word. However, the proposed approach still has the enhancement scope by optimizing the clustering algorithm. This chapter enhanced the text

representation by incorporating the feature clusters, but it did not include the advanced clustering algorithms. Therefore improving the clustering algorithm and learning embedding for phrases is the future direction of this work.

Chapter 6 TOPIC MODELING BASED FEATURE EXTRACTION FOR TIME SERIES PREDICTION

Feature extraction is gaining popularity in the field of ML and pattern recognition since the need for informative features is critical in extracting information and detecting patterns. The simplest time series prediction models use only information on the variable to be predicted and make no attempt to discover the factors that affect its behavior. Feature extraction is often useful in time series forecasting. In terms of public health, feature extraction from social media platforms has been crucial in delivering and getting information about any ongoing incident. This chapter demonstrates how feature extraction from a probabilistic model, such as the state-of-the-art LDA-based models, may offer a major basis for time series prediction utilizing online news articles and disease data concerning coronavirus disease. Section 6.1 reveals the overview of the entire chapter. Section 6.2 gives the details of the LDA model. Section 6.3 describes the proposed model, PAN-LDA. Section 6.4 discusses the experiments and results. Finally, Section 6.5 ends the chapter with the chapter summary.

6.1. Overview

Several years ago, the biggest challenge was finding information. Nowadays, a significant proportion of human knowledge and behavior is digitized on the web, namely in scientific papers, video, audio, images, emails, blogs, books, and social networks, addressing how to involve with these electronic records each time even more proficiently. The rapid development of advanced text mining tools in recent years has resulted in a substantial change in information extraction and prediction research. These strategies enhance present research efforts by increasing the efficiency and speed of existing methods. Since the advent of text mining technologies, there has been a greater emphasis on the study of extracting information from unstructured textual data.

One common approach for text mining is using the statistical model, such as topic models, to comprise enormous textual information based on the topic distribution of documents [309,310]. With the emergence of text mining methods, the research on topic modeling usually focused on textual data to obtain the results. While working with the unstructured textual data, there are numerous instances where the topic models, especially the LDA model, were utilized or improved to discover topics from the online text [311–313]. All these studies incorporate textual data for the evolution of topics in different ways. Although the traditional latent Dirichlet allocation model has been studied well, these methods do not consider numerical data. Recent studies have employed topic modeling methodologies to anticipate prices using unstructured data such as broadcast news and social media data [314]. In addition to online text mining for time-series predictions, researchers have widely applied sentiment analysis for processing

unstructured text. For example, X. Li, W.Shang, and S.Wang have considered news sentiment text features and grouped them according to their topics using topic modeling to increase the prediction accuracy [315,316].

While addressing unstructured textual data, one approach is to develop specialized search engines. For example, several researchers developed search engines to find the data related to a particular interest from the COVID-19-related publications across scientific disciplines [317–319]. However, such engines are limited to numerical data, keeping the textual data aside. Some studies focused on the textual information leaving behind the numerical data. For example, A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, et al. [320] used only the news article headlines for text mining. Another study mined only the new articles and forecasted Argentine and Brazilian currency markets' movement, which employed topic clustering, sentiment analysis, and regression analysis [321].

The availability of a massive amount of everyday data on online platforms has built a relationship between ongoing events and online data. Therefore, reducing the online textual data into topic distribution increase the value of this relationship. The numerical data can also be extended for technical analysis to extract more valuable information about the event with time.

However, to the best of our knowledge, none have defined a topic model based upon the combination of statistical data and textual data for prediction. Therefore, depending on which aspect is used, there is a scope for improvement. Our approach is to incorporate the time series numerical data along with online textual data in the standard LDA to extract better topics. Accordingly, we introduced PAN-LDA, a modified LDA, to take COVID-19 cases' numerical data into account to improve the feature extraction. The topic features are then presented in machine learning algorithms to have an advantage from our PAN-LDA model.

Various statistical methods have been developed, including ML techniques and time series methods, to track and predict the events' evolution with time. Regardless of the ongoing re-rise and prevalence of conventional ML algorithms, boosting strategies are still increasingly valuable for a medium dataset as the preparation time is generally extremely quick, and they do not require quite a while to tune its parameters. In recent years, gradient boosting-based ML methods such as Extreme Gradient Boosting (XGBoost) [322] and Light Gradient Boosting Machines (LightGBM) [323] have been applied by some researchers for a robust prediction of future events in different research fields [324–327].

Generally, the main components in achieving the outcome in the data mining task are data collection, data preparation, modeling, and evaluation. In the data preparation process, feature extraction models like LDA provide new lower dimension features. Accordingly, PAN-LDA is a topic model for extracting the features during the data preparation stage. Therefore, our focus will be on describing the flow of this model in various phases of data mining and performing experiments to understand the future benefits of

the extracted features. In the modeling phase, the PAN-LDA model's performance is evaluated using ML algorithms.

6.2. Latent Dirichlet allocation (LDA)

The most well-known example of topic modeling is Latent Dirichlet Allocation (LDA), a generative probabilistic model proposed by Blei et al. [138] to compute the latent topics from various text documents. Because of its great modularity, it can be quickly extended, giving much interest to its study. As a topic model, LDA aims to discover topic information in massive archives and may be extended to recognize patterns in generic data, images, social networks, and bioinformatics.

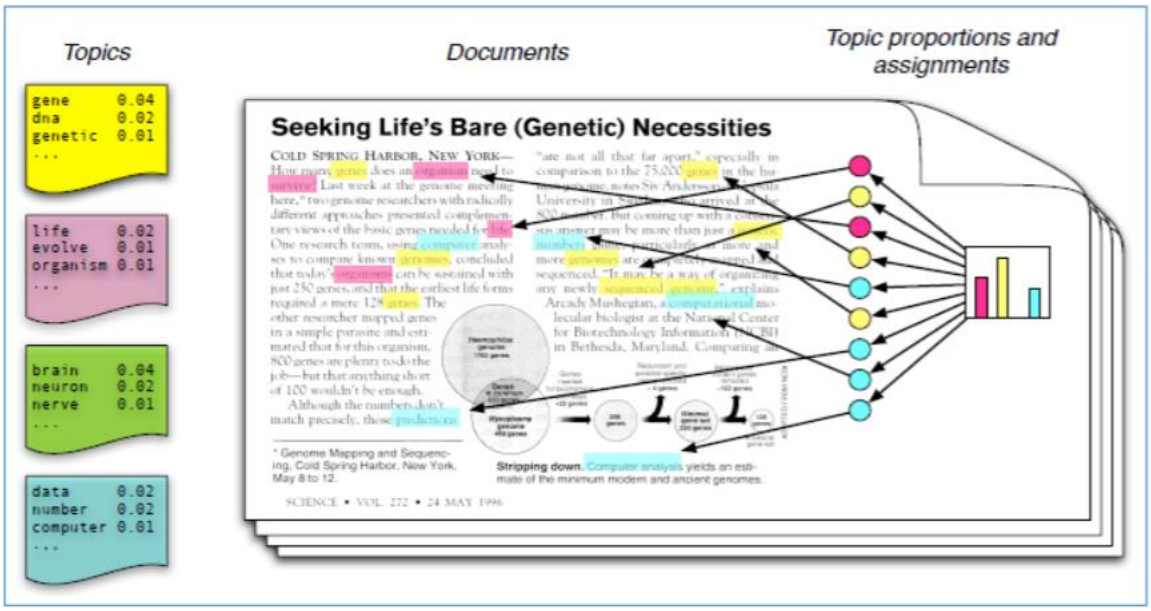


Figure 6.1 Generative Process for LDA [328]

Topic Modeling requires no prior annotations or data labeling since topic models can detect patterns in an unstructured set of documents and arrange these data at a level that human annotation cannot. Topic models as statistical models may be traced back to Hofmann's Probabilistic Latent Semantic Indexing (PLSI) in 1999 [329]. LDA, an improvement on PLSI, is perhaps the most often used topic model today. LDA aims to find brief descriptions of the subjects in the collection, condensing the profusion of original information into a shorter interpretable space while keeping the essential statistical connections to assist in successful data processing. In the case of text corpora, LDA assumes that each document includes several topics [330], with the topic probabilities providing a comprehensive explanation of the corpus.

LDA assumes that all documents in the collection have the same set of topics but that each document displays those topics in a variable proportion. Figure 6.1 depicts Blei's representation of the latent topic in a research article, which helps us better understand LDA's goal:

Figure 6.1 highlights words from various topics in different colors. For example, blue words represent data analysis, green words represent neurons, pink about evolutionary life, and yellow represents genetics keywords, indicating the content of the article and the model's understanding of a blend of several topics. LDA considers data observations resulting from a generative probabilistic process with latent variables. Furthermore, the word distribution of each topic is provided on the left, along with each word's appearance probability in each topic.

This method implies that the order of terms is not always significant. The same terms can be found in several topics with varying probability. Figure 6.1 shows how documents are modeled in LDA by considering that the topic distribution of all documents and the word distribution of all topics is known. With this information, the model then selects a topic from the topic distribution and then a word from the word distribution of that topic, continuing this process until the text is formed. This process is done for each document in the corpus. This procedure gives the LDA model its generative aspect, given certain hidden variables.

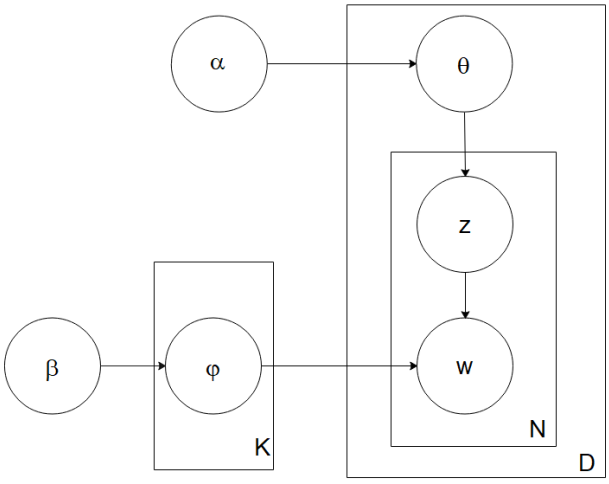


Figure 6.2 Graphical Model Representation of LDA

The generative process of LDA can be seen in Algorithm 6.1:

Algorithm 6.1: Latent Dirichlet Algorithm

$\forall\, k \in \{1, \ldots\ldots, K\}$

Simulate $\varphi_k \sim Dir(\beta)$

$\forall\, d \in \{1, \ldots\ldots, D\}$

Simulate $\theta_d \sim Dir(\alpha)$

$\forall\, w \in \{1, \ldots\ldots, N_d\}$

Simulate $z_{d,n} \sim Multi\ (\theta_d)$

Simulate $w_{d,n} \sim Multi\ (\varphi_{z_{d,n}})$

Parameters and variables:

- K: the total number of latent topics
- D: number of documents
- N_d: number of work tokens in document d
- α, β: Dirichlet parameters
- φ_k: per topic word distribution
- θ_d: per-document topic distribution
- $z_{d,n}$: topic index of the word
- $w_{d,n}$: index of the word w in document d
- $Multi$: Multinomial distribution

Figure 6.2 represents the graphical model of smoothed LDA. The $w_{d,n}$, the index of word w in document d, represents the input of the model. The output from the model is the K, predefined number of latent topics. Each topic k, $k \in \{1, \ldots\ldots, K\}$ is represented by a discrete probability distribution φ_k over the vocabulary V and generated from a Dirichlet distribution $\varphi_k \sim Dir\ (\beta)$. Additionally, every document d, $d \in \{1, \ldots\ldots, D\}$ comes from a Dirichlet distribution $\theta_d \sim Dir\ (\alpha)$, which is the topic distribution for each document d. From θ_d we calculate, $z_{d,n}$, per word topic assignment in document d, where β and α are the Dirichlet parameters.

6.2.1. Model Inference

Documents in LDA are represented as a combination of topics, with each topic having a certain probability of generating various words. As a result, LDA implies that data is generated through a generative process containing latent variables, which generates a combined probability distribution over the observed and hidden random variables, conveying the inferential problem to calculate the posterior distribution of the hidden variables given the observed variables as follows:

$$p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)} \quad (6.1)$$

For a set of given observed words $w = \{w_{d,n}\}$, inference methods aim to determine the posterior distribution over the unknown parameters. The posterior distribution cannot be calculated directly. As a result, approximation inference procedures are required to create an alternative distribution that is near to the posterior. There are several approximate inference algorithms, such as expectation propagation, Markov random fields, variational Bayes, Markov Chain Monte Carlo, etc. However, CGS is an effective inference technique to learn the model from data [331], where only the latent variable z is sampled, and the random variables such as φ and θ are marginalized out. Once the latent variables are sampled out, the random variables φ and θ can be estimated.

6.3. Proposed Model, PAN-LDA

This section describes our model PAN-LDA, followed by parameter inference. This model incorporates the changes in the number of daily new COVID-19 confirmed cases on a one-day interval along with the globally published news articles. Our model can discover hidden topics discussed in news articles related to COVID-19 in various countries.

6.3.1. Framework of the PAN-LDA model

Here, we discuss how the calculations flow when our model is used for COVID-19 case prediction (see Figure 6.3). The figure shows the flowchart representation of our proposed approach.

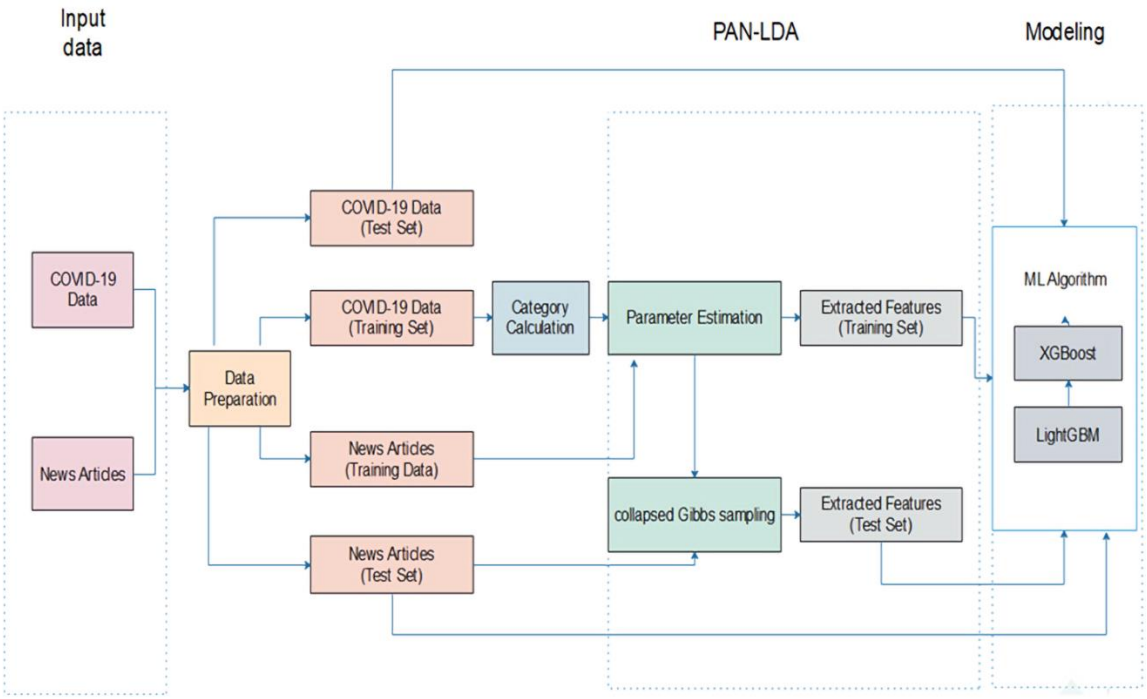


Figure 6.3 Flowchart of the Proposed Model, PAN-LDA

This chapter deals with the application of our PAN-LDA model for forecasting coronavirus cases over time. Firstly, a usual requirement in the data preparation phase is preprocessing the raw data. Our experiments preprocessed the collected text by noise removal, case folding, tokenization, stemming, lemmatization, and stopword removal. Next, in the data preparation phase, we focused on feature extraction while working on topic modeling with PAN-LDA.

In the data preparation phase, our model needs the word vectors and the statistics of the COVID-19 cases. The vector of words in PAN-LDA is provided by the 'bag-of-words 'model. However, the case statistics need to be calculated and categorized for our model PAN-LDA-see Figure 6.3. Our model generates three probability lists after training from the training set, i.e., topic distribution per document, word distribution per topic, and distribution of change in corona infected cases per topic. The topic distribution obtained serves as an additional feature for training an ML algorithm. The word distribution becomes the input parameter for inferring the topic from a previously unseen document, which as a result, generates the new topic distribution for the new document, which serves as an additional feature for the ML algorithm for predicting outcomes in the next phase of time series data.

In the modeling phase, we trained machine learning algorithms to compute regressions. In this chapter, we choose two machine learning algorithms, i.e., XGBoost and LightGBM, to perform our experiment.

The topic distributions from the training data, along with other features, were used to train the selected ML algorithms. Ultimately, the model can be utilized to forecast the selected index.

6.3.2. Model Description

It is noted that with the increase in the severity of pandemic cases, there is an association between topic generation and trend prediction. Inspired by this, we developed PAN-LDA, a modification of LDA, which incorporates the statistics of daily new coronavirus cases along with news articles for feature extraction.

The overall framework of the PAN-LDA approach is depicted in Figure 6.4. Our model incorporates the change in the number of reported coronavirus cases after a news article d is published, nc_d . Furthermore, the distribution of changes in reported cases per topic δ_k, is added in the figure to connect with the other distribution via nc_d . The past data of the infected corona cases are processed to find the change in the number of newly reported cases, $nc_{d,}$ after publishing the document d. The time lag that we considered for the collection of data is of one day. In the training process, data for daily new corona-infected cases were available. The presence of corona case data in PAN-LDA affects the distribution of changes in reported cases, affects the per-document topic distribution, θ_d and also the per topic word distribution, φ_k. After the parameter estimation, the latent topics of a document can be obtained. For a new document, the latent topic distribution is obtained using the estimated word distribution from parameter estimation in the inference methods on the document. The received topic features can then be introduced as input features in any ML method. In summary, the proposed model incorporates news articles and changes in the daily new corona infected cases to refine the parameter estimation and topic distribution inference on previously unseen documents. The obtained topic distribution can serve as input features for ML models to predict the time series.

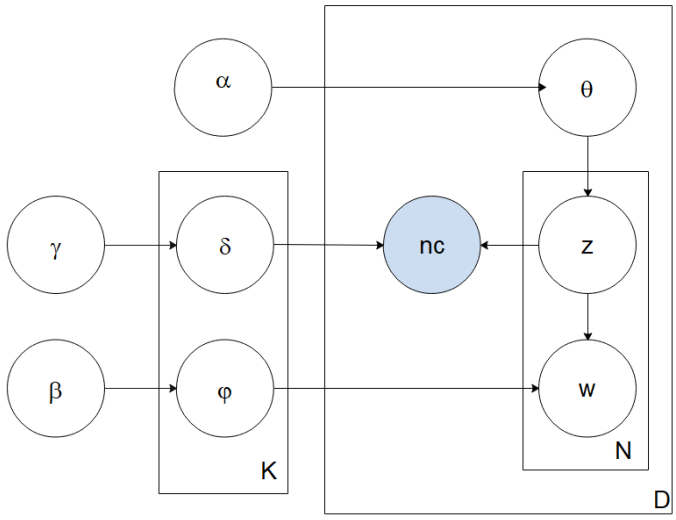


Figure 6.4 Graphical Model Representation of PAN-LDA

For a given collection of D documents having fixed vocabulary V, with N_d word tokens, $(w_{d,1}, \ldots\ldots, w_{d,N_d})$ in document d, each having an index in the vocabulary, $w_{d,n} \in \{1, \ldots\ldots, V\}$. It is assumed that the number of latent topics, K, is predetermined.

In LDA, per word topic assignment, $z_{d,n}$, is drawn from the probability list of K topics and depends on the previously drawn topic proportion in document d, $\theta_{d,k}$. And, a word instance, $w_{d,n,}$ is presumed to be deduced from a probability list of V words and depends on the word distribution, $\varphi_{d,k}$ and the topic index of the word, $z_{d,n}$. Based on LDA, PAN-LDA is also a probabilistic generative model. Accordingly, in PAN-LDA, the change in the number of new coronavirus disease infected confirmed cases, nc_d, is drawn from a list of probabilities of C categories and depends on the changes in the number of cases distribution δ_{z_d}.

Algorithm 6.2 describes the generative process of the PAN-LDA model:

Algorithm 6.2: PAN-LDA

$\forall\, k \in \{1, \ldots\ldots, K\}$

Simulate $\varphi_k\ Dir(\beta)$

Simulate δ_k ~$Dir(\gamma)$

$\forall\, d \in \{1, \ldots\ldots, D\}$

Simulate θ_d ~$Dir(\alpha)$

$\forall w \in \{1,\ldots\ldots,N_d\}$

Simulate $z_{d,n}$ ~$Multi$ (θ_d)

Simulate $w_{d,n}$~$Multi$ $(\varphi_{z_{d,n}})$

Simulate nc_d ~$Multi$ (δ_{z_d})

Parameters and variables:

- K: the total number of latent topics
- D: number of documents
- N_d: number of work tokens in document d
- α, β, γ: Dirichlet parameters
- φ_k: per topic word distribution
- δ_k: change in the number of daily COVID-19 cases distribution for topic k
- θ_d: per-document topic distribution
- $z_{d,n}$: topic index of the word
- $w_{d,n}$: index of the word w in document d
- $Multi$: Multinomial distribution

The variables are distributed via probability distribution:

$$p\,(z_{d,n} = k|\,\theta_d) = (\theta_d)_k \tag{6.2}$$

$$p(w_{d,n} = v \mid z_{d,n}, \varphi_1, \ldots\ldots, \varphi_K) = (\varphi z_{d,n})_v \tag{6.3}$$

$$p(nc_d = c|z_d\,, \delta_1, \ldots\ldots, \delta_K\,) = (\delta_{z_d})_c \tag{6.4}$$

The joint distribution of latent variables and observed data is then:

$$p(w, z, nc, \theta, \varphi, \delta|\alpha, \beta, \gamma) =$$
$$\prod_k p(\varphi_k|\beta) \prod_k p(\square_k|\,\gamma) \prod_d [\,p(\theta_d|\,\alpha)[\,\prod_n p(z_{d,n}|\,\theta_d)\,p(w_{d,n}\mid z_{d,n}\,, \varphi)]p(nc_d|z_d, \delta)] \tag{6.5}$$

Inferences about the posterior parameters typically yield topics where the probability mass of each topic is assigned to frequently co-occurred words that are semantically strongly related.

6.3.3. Topic Inference

The central issue in topic modeling is posterior inference, which includes learning the posterior probabilities of the observed data, i.e., words in documents, w, and the change in the number of reported corona infected cases after the documents were published, nc, and the latent variables, i.e., θ, φ, δ and z. In PAN-LDA, the posterior probabilities of latent variables can be calculated as :

$$p(\varphi, \theta, \delta, z|w, nc, \alpha, \beta, \gamma) = \frac{p(\varphi,\theta,\delta,z,w,nc,\alpha,\beta,\gamma)}{p(w,nc,\alpha,\beta,\gamma)} \quad (6.6)$$

Unfortunately, the computation of this posterior distribution is intractable. The computation of the normalization factor, particularly $p(w, nc, \alpha, \beta, \gamma)$, cannot be done accurately. Among the various inference methods, collapsed Gibbs sampling (CGS) proposed by T. Griffiths and M. Steyvers [331] is known for model estimation with high accuracy.

Using the CGS with LDA, we are interested in computing the posterior distribution of a topic z is allocated to a word w, given the remaining words are assigned to other topics, as follows:

$$p(z_i \,|z_{-i}\,, w, \alpha, \beta, \gamma) \quad (6.7)$$

where z_{-i} means all topic allocations, excluding z_i.

Also, we must appeal to approximated inference, where some of the parameters are marginalized out. Therefore, we applied collapsed Gibbs Sampling for finding inference, which marginalizes out parameters φ, θ, and δ and, on each iteration, recovers the topic, $z_{d,n}$ of a word token w, from a distribution conditioned on the present values of remaining variables. In a space containing all the variables, by sampling in a collapsed space, collapsed Gibbs Sampling usually converges much faster than a common Gibbs sampler. By simply computing all the K topic assignments, a naive implementation of equation (6.8) has a complexity of O(K) per token. The posterior distribution for sampling the latent variable z, after the random variables are marginalized out, for PAN-LDA, is:

$$p(z_{d,n} = k \big| z_{-d,n}, w, nc, \alpha, \beta, \gamma) \propto (N_{d,k} + \alpha) \frac{\left(N_{k,w_{d,n}}^{-d,n} + \beta\right)}{(N_k^{-d,n} + V\beta)} \frac{\left(N_{k,nc_d}^{-d,n} + \gamma\right)}{(N_k^{-d,n} + C\gamma)} \quad (6.8)$$

where,

$N_{d,k}$ denotes the total words allocated to topic k in document d.

$N_{k,w_{d,n}}$ is the number of words of type $w_{d,n}$ allocated to topic k.

N_k is the total words allocated to topic k.s

N_{k,nc_d} is the number of words belonging to category nc_d and assigned to topic k.

For topic modeling, we estimated the topic distribution per document θ_d, the word distribution per topic, φ_k, and the topic assignments per word $z_{d,n}$. Using topic allocations, φ, θ, and δ can be computed as:

$$\theta_{d,k} = \frac{N_{d,k}+\alpha}{N_d+K\alpha} \tag{6.9}$$

$$\varphi_{k,w_{d,n}} = \frac{N_{k,w_{d,n}}+\beta}{N_k+V\beta} \tag{6.10}$$

$$\delta_{k,nc_d} = \frac{N_{k,nc_d}+\gamma}{N_k+C\gamma} \tag{6.11}$$

Equations (6.9)-(6.11) estimate the quantity probabilities that the word w belongs to topic k, $\varphi_{k,w_{d,n}}$, topic k is generated in document d, $\theta_{d,k}$ and the document d with category nc_d assigned to topic k, δ_{k,nc_d}, respectively. The CGS procedure is outlined in Algorithm 6.3.

Algorithm 6.3: collapsed Gibbs sampling

Input: $w \in d, nc$

initialize z and increment counters
for each iteration do
    for each document do
        for n = 1 to N_d do
            topic = $z[n]$
            decrease $N_{d,topic}$, $N_{topic,w}$, $N_{topic,nc}$ and N_{topic} by one
            for each topic k do

$$p(z_{d,n} = k | z_{-d,n}, w, nc, \alpha, \beta, \gamma) \propto (N_{d,k} + \alpha) \frac{\left(N_{k,w_{d,n}}^{-d,n} + \beta\right)}{(N_k^{-d,n} + V\beta)} \frac{\left(N_{k,nc_d}^{-d,n} + \gamma\right)}{(N_k^{-d,n} + C\gamma)}$$

            topic ← $\sim p(z| \cdot)$
            $z[n]$ ← topic
            Increase $N_{d,topic}$, $N_{topic,w}$, $N_{topic,nc}$ and N_{topic} by one
return z, $N_{d,k}$, $N_{k,w}$, $N_{k,nc}$ and N_k

where count matrices are the following notations: $N_{d,k}$, $N_{k,w}$, $N_{k,nc}$ and N_k .

As shown in Figure 6.4, other parameters of the model will not get affected by the δ and γ , if there is an unavailability of the corona case data. Therefore the inference from a previously unseen document, without having historical case data, will be the same as the inference from LDA.

The following pseudo-code, i.e., Algorithm 6.4, shows the inference from a new document without θ, φ, and δ.

Algorithm 6.4: Topic Inference from a Previously Unseen Document

Input: w_{pu} and φ

initialize α, β

initialize $z_{pu,n}$

for each iteration do

for i = 0 to N_{pu-1} do

topic ← $\sim p(z_{pu,n} = k | z_{-pu,n}, w_{pu}, \alpha, \beta, \varphi) \propto \frac{\left(N_{pu,k}^{-pu,n} + \alpha\right)(N_{k,w_{new,n}}^{-pu,n} + \beta)}{(N_{k}^{-pu,n} + V\beta)}$

update $\theta_{pu,k}$

end

end

6.4. Experiments and Results

This section discusses the experimental setup and the comparative results. As discussed, we experimented with four different models and explored four different feature sets, i.e.:

- COVID-19 cases statistics as a base feature, FS1
- COVID-19 cases statistics with topic distributions from LDA, FS2
- COVID-19 cases statistics with topic distributions from LDA and sentiment scores to the latent topics, FS3
- COVID-19 cases statistics with topic distributions from PAN-LDA, FS4

In FS1, only historical COVID-19 daily case data are used as base features in the prediction models. For FS2, in addition to the historical data, the topic distribution from the LDA is integrated into the prediction model. In FS3, we use the COVID-19 cases data along with the computed topic distribution of the news articles and sentiment scores specific to the extracted latent topics. Since the topics extracted from news

articles do not have any associated sentiments, sentiment analysis of reviews is also done by using VADER to compute the relative sentiment scores with respect to the topics. The historical data, extracted topics, and their sentiments are used as input features to a Machine Learning prediction model. Finally, FS4 denotes the feature set obtained from the topic distributions from PAN-LDA and the COVID-19 historical data.

Once the models were trained, we used the backtesting approach for time series forecasting. For backtesting, we used the walk-forward testing [332] routine, which accounts for model performance at different time windows.

6.4.1. Data Selection and Gathering

COVID-19 data: The data for the number of confirmed coronavirus infected cases used to experiment are the official data published by 'Our World in Data' [333]. They have provided global and reliable data to study statistics on the COVID-19 pandemic. The dataset is updated daily from the WHO situation reports [334]. We use available data on the daily new cases of coronavirus-infected people from January 2020 to May 2020.

News Articles: The news articles dataset used in this chapter was gathered from Aylien [335]. Aylien has aggregated and published the COVID-19 dataset that can be used to analyze global news throughout the outbreak. Aylien has transformed the COVID-19 dataset into structured and actionable data using NLP and ML. The data analyzed in this study correspond to the period that stretches from January 2020 to May 2020. As a result, a total of 1147454 articles were considered for the experiment.

6.4.2. Data Preparation

Consequently, we trained our model on a collection of more than 1 million news articles, which contributed to text documents for this experiment. We preprocessed the data by tokenization, stemming, and removing stop words. The text in new articles is represented by a vector using the 'bag-of-words' model in Gensim [336].

We collected one-day level new corona infected data and classified the changes in the number of infected cases into three different categories. The threshold was considered based on an average of one-day change for collected data. There is an average change of 0.1285% in both directions. This chapter experimented with a 0.10% - 0.15% change as the threshold. As a result, the data is divided into three categories, i.e.,

- category 1, if the change in the number of new coronavirus cases lies above the threshold,
- category -1 if the change in the number of coronavirus cases is below the threshold, and

- category 0 if it lies within the threshold

So, the value for nc_d in PAN-LDA is explained in three levels as below :

$$nc_d = \begin{cases} 1, if\ \frac{number\ of\ the\ new\ coronavirus\ case_{t+1\ day} - number\ of\ the\ new\ coronavirus\ case_t}{number\ of\ the\ infected\ person_t} * 100 > 0.15 \\ -1, if\ \frac{number\ of\ the\ new\ coronavirus\ case_{t+1\ day} - number\ of\ the\ new\ coronavirus\ case_t}{number\ of\ the\ infected\ person_t} * 100 < 0.1 \\ 0, otherwise \end{cases} \quad (6.12)$$

We found that 767345 articles fell in category -1(nc_d =-1), 145773 articles in category 0(nc_d = 0) and category 1(nc_d =1) has 234336 articles.

After all of the data has been preprocessed, we then divide it into training and test sets using walk-forward validation [332], a variant of cross-validation. The whole dataset was split into 11 overlapping datasets with a 19-day window. Each of these datasets was divided into training and test sets as 25% testing and 75% training.

The input dataset for PAN-LDA consists of all the word tokens in D documents, $w_{d,n}$ and category values, nc_d. Next, we set the hyperparameters of the Dirichlet distributions, i.e., α, γ, and β. In order to have a few words with a high probability per topic and numerous latent topics with a high probability per document, the values of Dirichlet distributions were set as $1/K$,i.e., $\alpha = \beta = \gamma = 1/K$ for all the topic models.

Selecting the optimum number of topics (K) in topic modeling is also a significant problem. In order to estimate the optimum values of K and the number of iterations (iter), we ran 200 iterations, iter=200, and noted the computed value at every 20 iterations. We evaluated the effect of iteration count with different numbers of topics on log-likehoods at these savings points. The results are presented in Figure 6.5, which depicts the stability in results when iter > 140.

Also, to better understand the parameters' values, we computed the log-likehoods for PAN-LDA in Figure 6.6. The graphs in Figure 6.6 suggest that the optimum value is achieved at K=10. And as the graph showed stable results when iter > 140, we set the iter = 160 for our experiment. Accordingly, we set the same values for the LDA model.

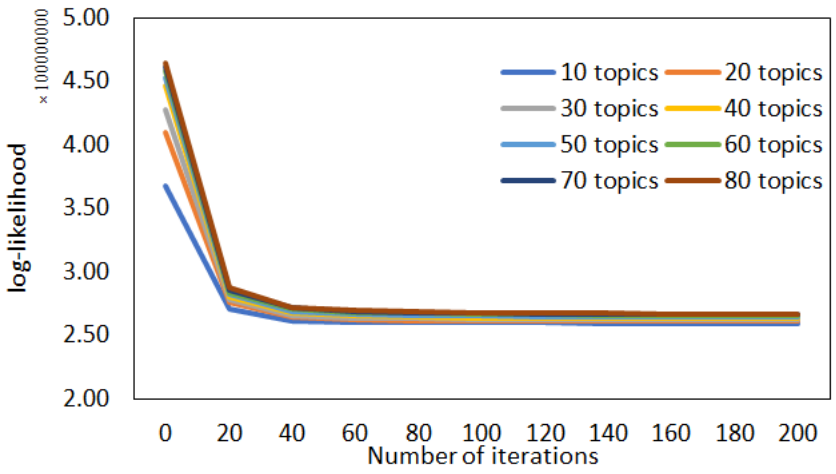
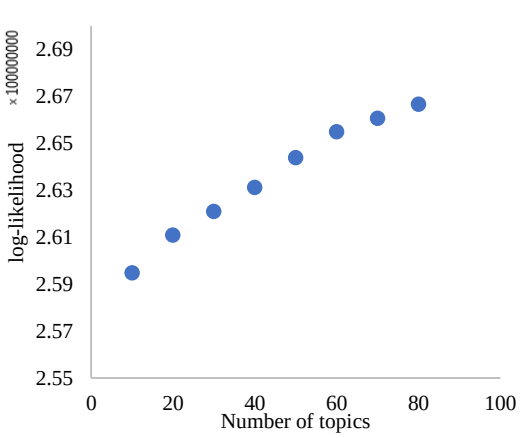
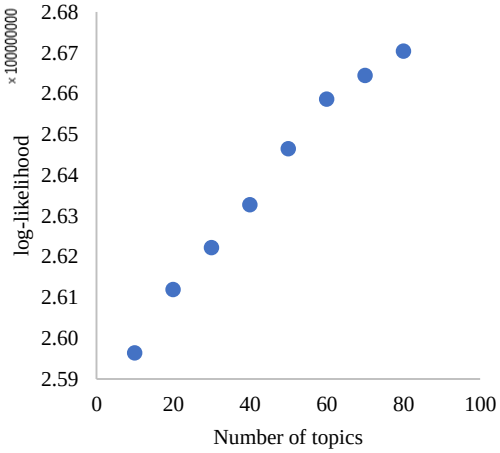


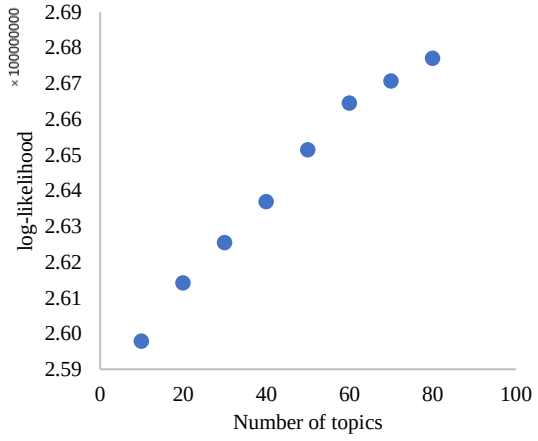
Figure 6.5 The log-likelihood for PAN-LDA with collapsed Gibbs sampling



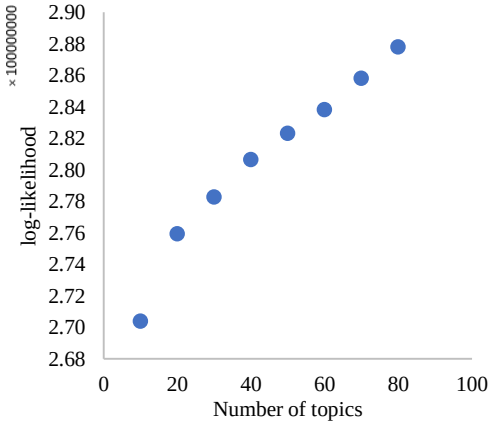
(a) 200 iterations



(b) 140 iterations



(c) 100 iterations



(d) 20 iterations

Figure 6.6 The log-likelihood Against the Number of Topics

After setting the parameter values, we extracted topics from the LDA, news-text-sentiment feature grouping using LDA and PAN-LDA models. Table 6.1 shows 6 of the 10 topics discovered by LDA and PAN-LDA with their top 15 words.

Table 6.1 Examples of Topics Generated by LDA and PAN-LDA

Sports	Finance	Business	Entertainment	Country	Politics
League	market	business	time	australia	trump
Football	economy	company	people	reuters	president
Season	global	ceo	lockdown	new_zealand	donald_trump
Players	china	officer	social_media	government	white_house
Club	markets	financial	instagram	australian	president_trump
premier_league	oil	year	family	bank	house
Team	economic	bank	star	editing	virus
Sports	year	industry	quarantine	reporting	washington
England	energy	companies	year	reuters_reuters	bill
Training	reuters	airline	video	sydney	congress
United	virus	insurance	food	european	year
Sport	stock	group	life	germany	trump_administration
Games	bank	chief_executive	social	france	fox_news
Clubs	prices	cash	facebook	prime_minister	senate
Game	demand	businesses	times	french	federal

Sports	Finance	Business	Entertainment	Country	Health
League	bank	company	instagram	new_york	virus
Football	economy	business	time	city	hospital
Season	market	year	star	county	patients

Team	economic	ceo	video	california	infection
Players	global	officer	social_media	governor	disease
Sports	markets	industry	live	york	health
Game	financial	companies	twitter	florida	vaccine
Games	oil	sales	family	texas	testing
Club	reuters	market	music	los_angeles	symptoms
Events	stimulus	stock	film	virus	test
Time	energy	group	story	department	tests
Year	prices	production	years	order	people
premier_league	unemployment	supply_chain	life	mayor	fever
United	money	quarter	series	chicago	medical
Event	rate	nasdaq	netflix	people	medicine

Table 6.1 suggests that some topics from both models have the same words or words with similar implicit meanings. Though, their ranking order, suggesting their importance, is different. Moreover, some topics from the two models are entirely different.

Based on the extracted words, we interpreted the meaning of the topic and assigned labels to each, i.e., 'Sports', 'Finance', 'Business', 'Entertainment', 'Country', 'Health', and 'Politics'. We assign the same color to the words belonging to the same topic. Topic 1, 'Sports', colored in orange, has similar sets of words for both the models. Though the words in the models have a different order, indicating their importance. Topic 2, 'Finance', also contains similar words for both the models, yet, the absence of an important word, i.e., "stock", can be noted in the vocabulary words of PAN-LDA.

Similarly, the word 'unemployment' is absent from the top vocabulary words of LDA. In topic 3, 'Business', the words generated by both the models are quite different but appear to be similar in their implicit meanings. In the next topic, i.e., 'Entertainment', PAN-LDA has generated more meaningful words related to the topic than LDA. In LDA, the words are vaguely present and do not contribute much to extracting a single topic. The words from LDA in topic 5 seem to be a combination of two topics. PAN-LDA isolates more coherent topics, such as health, and social anxiety, as compared to LDA. In LDA, the remaining words talk about politics, lockdown, etc. We noted that the rest of the generated words in LDA do not contribute much to forming new identifiable topics. Also, the remaining topics

generated by PAN-LDA are vaguely present in LDA. Some topics in LDA seemed to be a combination of topics from PAN-LDA. Additionally, PAN-LDA produced more identifiable topics.

As the topics obtained from the models are mainly used for inferring the topic distributions from the text, the interpretation and meaning of topics are not of much concern in this experiment. From the parameter estimation process, the topic distributions for all training set documents along with estimated word distributions for inference were obtained. The obtained topic distribution then served as an input feature for ML algorithms without interpreting their meaning. It can be noted that two different models generated different topics. This suggests that adding a new feature, i.e., changes in data of new corona cases, successfully influenced per topic word distribution in the parameter estimation, which is evaluated in the next step.

Following that, using the estimated φ values and all words in the test set, $w_{d,n}$, the topic distributions were inferred from documents in the test set. The resulted topic distributions were fed into machine learning models, i.e., XGBoost and LightGBM, for testing in the next phase.

The prepared data is then arranged into four feature sets, FS1, FS2, FS3, and FS4, for both training and test sets.

6.4.3. Evaluation Indicators

We used four widely accepted statistical indicators, i.e., R^2, RMSE, MAE, and Mean Absolute Deviation (MAD), for performance comparison of ML algorithms trained with the above-said feature sets. Each metric is computed, as mentioned in Table 6.2.

Table 6.2 Performance Metrics and Their Calculations

Metrics	Calculation
R^2	$1-\frac{\sum(a_i-f_i)^2}{(a_i-\bar{a})^2}$
RMSE	$\sqrt{\frac{1}{N}\sum_{1}^{N}(a_i-f_i)^2}$
MAE	$\frac{1}{N}\sum_{1}^{N}\lvert a_i-f_i\rvert$
MAD	$\frac{1}{N}\sum_{1}^{N}\lvert a_i-\bar{a}\rvert$

where N corresponds to the total in the test data, a and f are the i^{th} value of the observed and forecasted number of new COVID-19 cases in the testing period and ā denotes the mean value of c.

It should also be noted that lower values of the RMSE, MAE and MAD indicate a better fit.

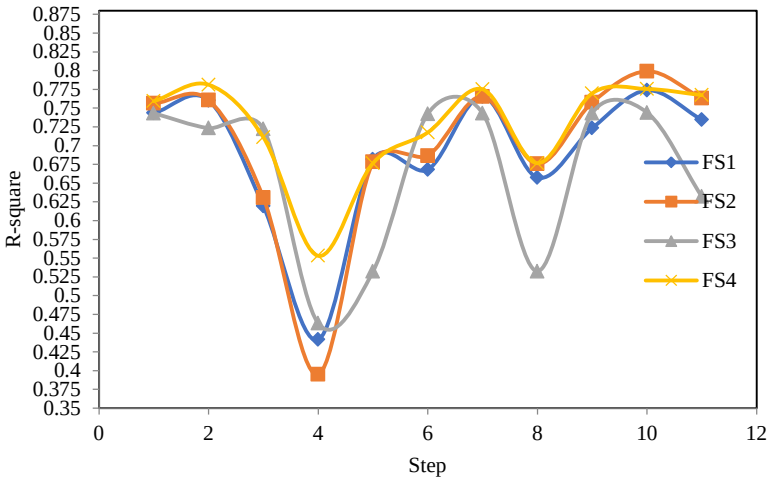
6.4.4. Results

Though the focus of our study is on extracting the better feature in the data preparation stage, yet we use two ML algorithms, i.e., XGBoost and LightGBM, to validate the performance of PAN-LDA. Next, we sought to evaluate the models using four statistical metrics, i.e., R^2, RMSE, and MAE provided by scikit-learn [337–339] and MAD provided by mad function in class Series in pandas [340].

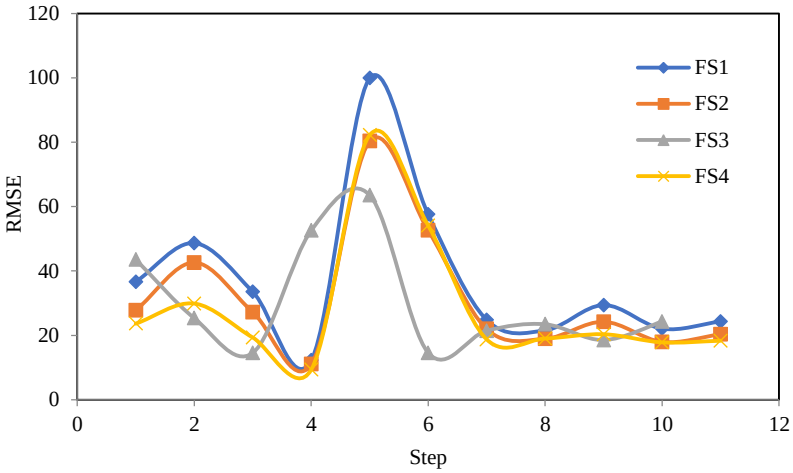
6.4.4.1. Results of XGBoost

Initially, as we were interested in a fair comparison among the results of different feature sets, we trained the XGBoost with the default parameter values [341] in the modeling phase.

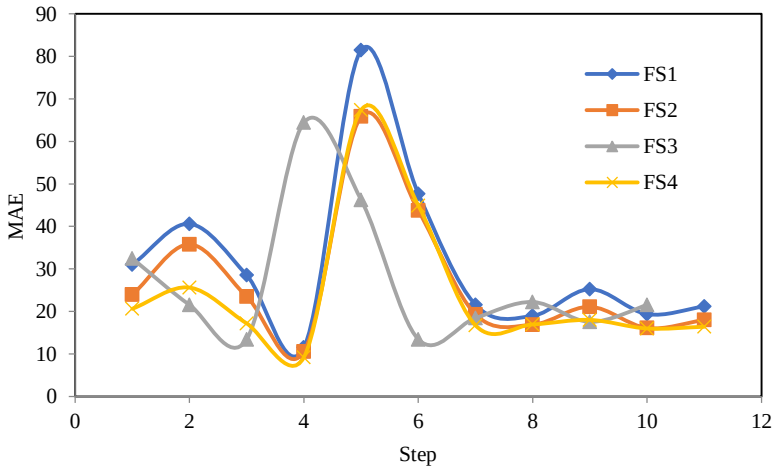
To better understand the XGBoost results, Figures 6.7(a)–(d) show the distribution of the four evaluation metrics mentioned in this chapter. In each figure, we compare the outputs (vertical axis) of evaluation indicators for all four feature sets against each step of walk-forward testing.



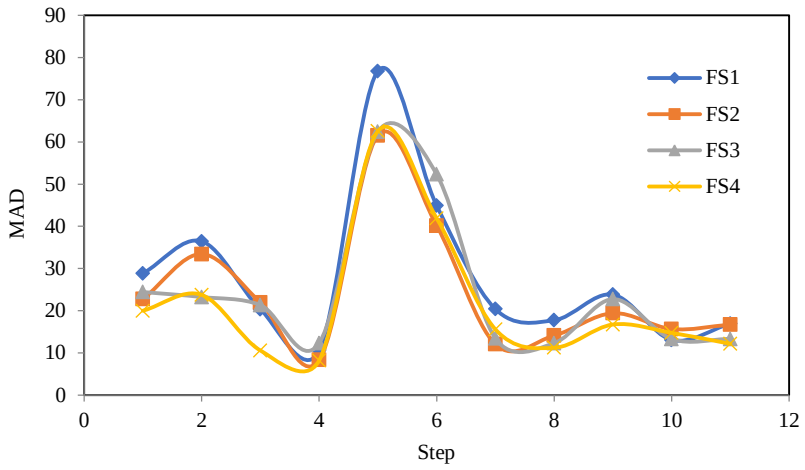
(a)


120
100
80
60
40
20
0
RMSE
0
2
4
6
8
10
12
Step
FS1
FS2
FS3
FS4

(b)


90
80
70
60
50
40
30
20
10
0
MAE
0
2
4
6
8
10
12
Step
FS1
FS2
FS3
FS4

(c)

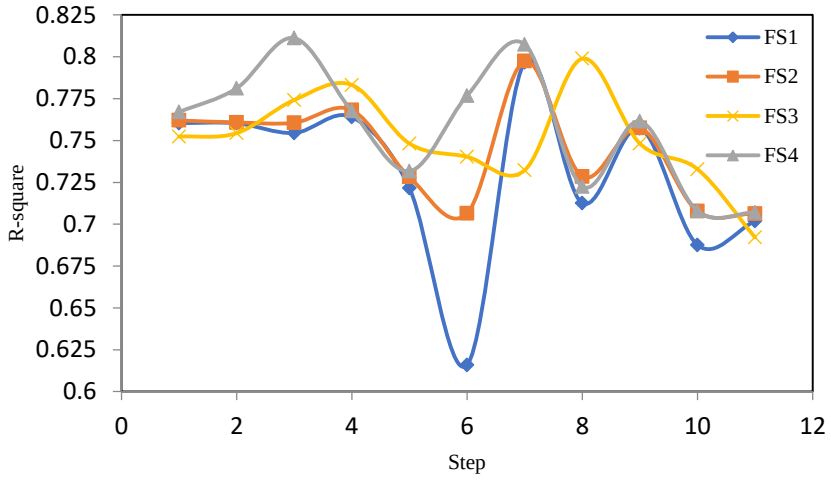


(d)

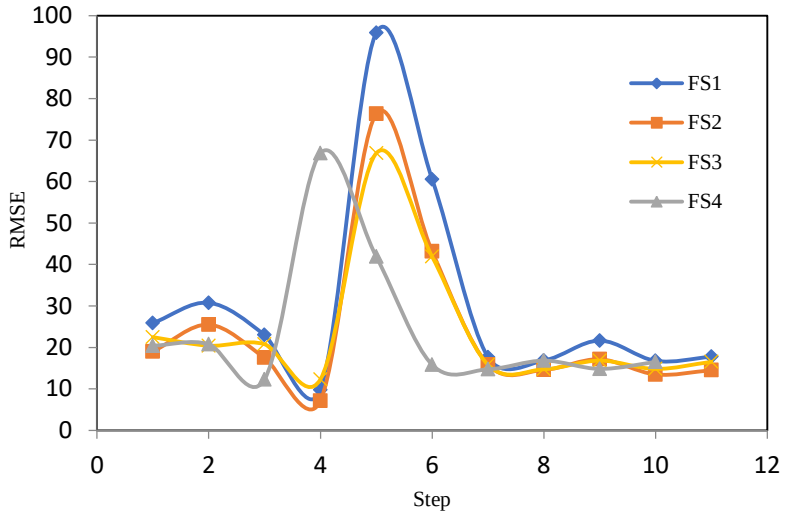
Figure 6.7 The presentation of (a) R^2, (b) RMSE, (c) MAE, and (d) MAD Between the Actual and the Predicted Number of New Confirmed Cases for FS1, FS2, FS3, and FS4 by XGBoost

6.4.4.2. Results of LightGBM

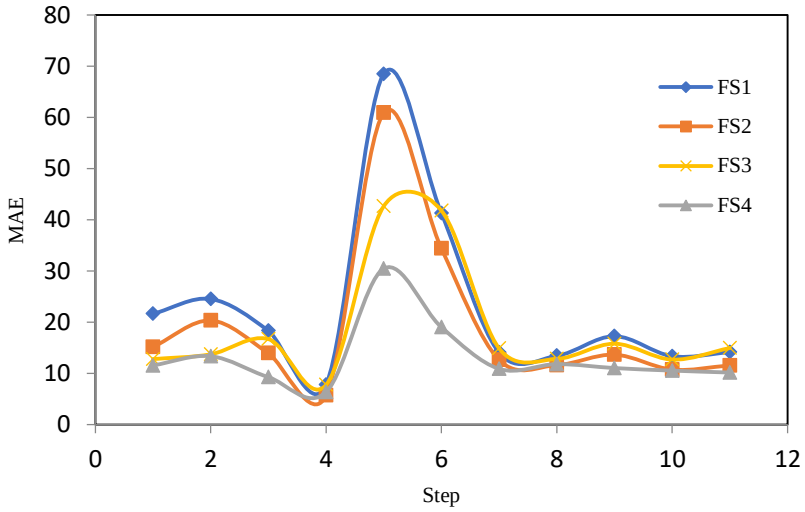
The results of all the feature sets from LightGBM are presented using 11 overlapping training–test sets from the walk-forward validation. Figures 6.8 (a)-(d) illustrate the performances of LightGBM with different sets of features, i.e., FS1-FS4.



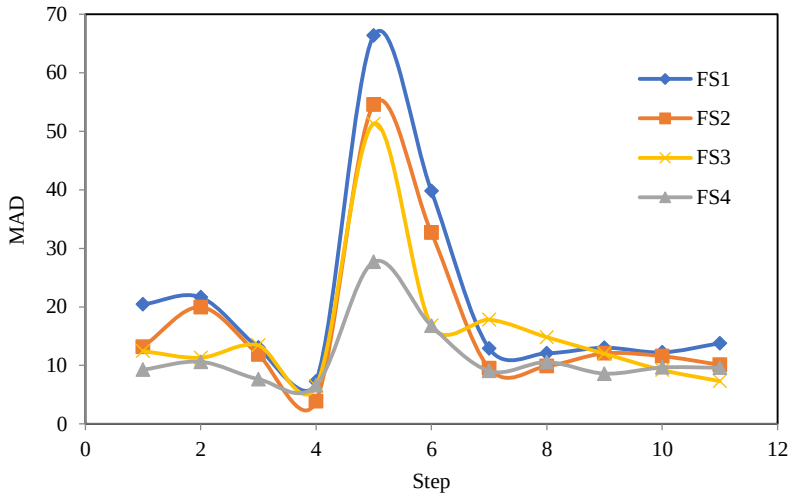
(a)



(b)



(c)



(d)

Figure 6.8 The presentation of (a) R^2, (b) RMSE, (c) MAE, and (d) MAD Between the Actual and the Predicted Number of New Confirmed Cases for FS1, FS2, FS3, and FS4 by LightGBM

Table 6.3 Comparison of Results of XGBoost and LightGBM with FS4

Step	R^2		RMSE		MAE		MAD	
	XGBoost	LightGBM	XGBoost	LightGBM	XGBoost	LightGBM	XGBoost	LightGBM
1	0.7592	**0.7670**	23.6395	**15.7655**	20.6263	**11.5822**	19.9906	**9.2664**
2	**0.7812**	**0.7812**	29.8801	**16.7653**	25.6069	**13.3801**	23.7570	**10.6228**
3	0.7112	**0.8112**	19.3234	**11.6423**	17.1817	**9.29156**	10.5774	**7.6562**
4	0.5531	**0.7676**	9.34649	**07.9876**	9.21931	**6.37479**	8.18504	**6.6132**
5	0.6766	**0.7317**	82.2345	**78.2345**	67.3902	**30.4953**	62.5700	**27.6917**
6	0.7175	**0.7768**	54.0757	**28.8765**	44.9171	**19.0459**	41.9074	**16.7886**
7	0.7752	**0.8075**	18.7480	**13.6653**	16.7225	**10.9061**	15.5794	**9.0570**
8	0.6766	**0.7226**	19.0169	**14.8265**	16.9371	**11.8328**	11.2149	**10.5660**
9	**0.7697**	0.7615	20.3450	**13.8643**	17.9971	**11.0649**	16.7289	**8.6037**
10	**0.7755**	0.7079	17.8170	**13.2345**	15.9795	**10.5622**	14.7570	**9.6441**
11	**0.7673**	0.7070	18.3331	**12.7654**	16.3914	**10.1879**	12.1495	**9.6112**
avg	0.7239	**0.7583**	28.4327	**20.6934**	24.4517	**13.1567**	21.5834	**11.4655**

After text mining became popular and viable for extracting information from text, public health research often incorporated unstructured textual data. This chapter presented a feature extraction model based on changes in daily COVID-19 cases data and news articles. To derive the inputs for the ML prediction models, we used the PAN-LDA model to extract relevant features. In order to take advantage of our PAN-LDA, the topic distributions from PAN-LDA are then used in a machine learning model. We compared the performance of our approach in predicting COVID-19 cases one day after news articles were released. By comparing the final results while employing the four different feature sets, FS1, FS2, FS3, and FS4, an experiment was conducted to demonstrate the benefits of integrating the features generated using PAN-LDA. Furthermore, we compare the results from XGBoost and LightGBM for all four feature sets, using the 11 overlapping training-test sets. The proposed model's prediction errors were lower than those of the other techniques. When the features from the proposed model are employed in both XGBoost and LightGBM, the results reveal that they empirically add value to the prediction.

The results for R^2, RMSE, MAE, and MAD for XGBoost are provided using graphs in Figure 6.7(a)-(d). We observed that the best results are obtained with FS4 for 7 out of 11 overlapping datasets with all statistical measures. Comparing the results with FS1, FS2, and FS3, as shown in Figure 6.7(a)-(d), shows that FS1, FS2, and FS3 have larger values for average RMSE, average MAE, and average MAD but smaller average R^2 than the FS4. Compared to the baseline methods, the proposed model improves average RMSE by 24-3 percent and MAE by 22-7 percent. The MAD in Figure 6.7(d) reveals that results from XGBoost when using FS2 were better in some datasets, but the best result for average MAD is achieved for the FS4, followed by FS2, FS3, and FS1. We concluded that FS4 provided better input features than FS1, FS2, and FS3. The result is consistent in all of the evaluation indicators. Also, the XGBoost gave the best average performance when used with the feature set FS4.

The results of LightGBM are shown in Figures 6.8 (a)-(d). These figures show that the FS4 has smaller RMSE, MAE, and MAD than FS3, FS2, and FS1 but larger R^2 in 8 out of 11 steps of backtesting. It implies that the performance of LightGBM when using features from PAN-LDA was better than that when using other sets of features. The average performance from 11 overlapping datasets indicates that LightGBM with FS1 has the worst performance, followed by FS2 and FS3, and the best with FS4 features. The R^2 for FS4 was improved by 3.84% than when using FS1. The average RMSE from PAN-LDA, 20.6934, is significantly better than FS1, 30.5913, FS2, 24.0235, and FS3, 27.1294. The average MAE and MAD show the same results. Figure 6.8(d) shows that the LightGBM, when used with the features from our model, outperformed FS1 by 45.78% and FS2 by 33.39%, and FS3 by 27.06%. Additionally, in Figures 6.8 (a)-(d), the performance of all the feature sets is compared by taking all the evaluation metrics, showing the benefit of PAN-LDA clearly. These figures also suggest that the

performance with FS4 was much better than FS1, FS2, and FS3 for all the evaluation metrics, namely R^2, RMSE, MAE, and MAD.

As for both XGBoost and LightGBM, on average, the results from FS4 are better than the results from FS3, FS2, and FS1. Figure 6.7(a)-(d) shows that the PAN-LDA model outperforms the baseline models in terms of MAE but not so much in terms of RMSE as RMSE penalizes larger prediction errors, while MAE stands for the absolute difference between observed and predicted values. Therefore, in Table 6.3, a comparison of these two machine learning algorithms using FS4 has been demonstrated. It can be noted from Table 6.3 that the highest correlations were achieved for LightGBM with an average R^2 of 0.7583. Also, the LightGBM has a smaller average value of RMSE, MAE, and MAD than XGBoost.

The experimental results and comparison of the proposed PAN-LDA model's performance with baseline models clearly show that supplementary/side information, such as new article content, is a valuable and expressive source of information for improving ML algorithms' predictions. Because historical data only captures the general perception of the target item, it cannot be used to generate precise forecasts. The features extracted from the LDA model do not seem to give much advantage for data forecasting over time. Moreover, incorporating sentiment scores as an additional feature in the prediction model has improved performance with less prediction error, such as MAE and RMSE. Besides, the results from FS1 are the worst for both XGBoost and LightGBM with the walk-forward testing. It can be concluded that incorporating the infectious disease data along with news articles in PAN-LDA gave better performance than LDA, which incorporates news articles only. This suggests the benefit of additional features in PAN-LDA. However, it seems that adding XGBoost resulted in only a few changes with the PAN-LDA model. Overall, it can be implied that LightGBM can forecast more closely to the actual values of COVID-19 cases than the XGBoost method. Also, including changes in the number of COVID-19 cases into account in PAN-LDA for prediction with time series, esp. with LightGBM.

In this study, the overall time period of the research is short because of the limited availability of reliable new article data [335]. We will improve our model with more data in the future.

6.5. Chapter Summary

In this work, we proposed an LDA-based mathematical model, PAN-LDA, which integrates news articles and data of confirmed COVID-19 cases for better feature extraction. The resultant features can be input as additional features to any ML algorithm to forecast trends with time series. Our chapter introduced the extracted features from the PAN-LDA model to two gradient boosting-based ML algorithms, i.e., XGBoost and LightGBM, to validate the feasibility of applying PAN-LDA compared to baseline methods. The features from PAN-LDA significantly added value to the goal output when used in ML

algorithms. Moreover, LightGBM gave a considerably better performance than XGBoost. In summary, the features from PAN-LDA generated more identifiable topics and empirically added value to the prediction when they were used in LightGBM.

Chapter 7 APPLICABILITY OF INTERNET PLATFORMS TO ENHANCE PREDICTION OF EPIDEMIC MODELS

Social media platforms are among the most extensively utilized communication channels and have become an integral part of our daily lives because of the speed and low cost these services give to their users. Social media plays a crucial role in notifying and informing the population during disease outbreaks. Raising knowledge of the disease and its prevention can lead to a change in behavior, which in turn influences contact/incidence rates. Web-based data may also be utilized to create, build, and parameterize models. As digital technology continues to advance and expand, online data is predicted to play an increasingly important role in infectious disease modeling in order to increase predictive power. Various information categories have the capacity to affect human ideas and decision-making; thus, it is critical to analyze numerous web-based platforms with varying information formats (such as comments, images, blogs, and short text) to acquire a comprehensive view. [342]. This chapter presents an epidemic model that incorporates the features from various online data sources for forecasting pandemic cases. Section 7.1 presents the overview of the chapter. Section 7.2 gives details of the susceptible-infected-removed (SIR) model, LSTM neural network, and PAN-LDA model. Section 7.3 provides the details of the proposed model that incorporates the data from online and social media platforms. Section 7.4 presents the results. Finally, section 7.5 concludes the summary of the entire chapter.

7.1. Overview

Social media have transformed the manner and speed with which information is disseminated. This paradigm change in the way information is transmitted has been used in various fields other than its familiar application in product marketing and promotion as a source of market intelligence and consumer involvement. The Centers for Disease Control and Prevention utilized Twitter to provide updates and promote flu prevention techniques to help curb the spread of H1N1 influenza, with its network growing throughout the outbreak [343]. Aside from encouraging behavior change, reporting symptoms and disease status published in social media postings may be beneficial in detecting and even forecasting the path of an epidemic. Researchers are increasingly interested in incorporating social media into mathematical models.

With the emergence of a large-scale outbreak of infectious disease and the declaration of a significant public health emergency, the public uses epidemiological models to estimate and anticipate the growth trend of the disease, and the results are used to drive the formation of preventative and control measures [203,344–346]. The most common classic epidemiological models are SI, SIR, and SEIR, where 'E' refers to the exposed population [347]. The SI model has two states: susceptible(S) and infectious(I), with the

change in the size of each state depending on predetermined differential equations. Considering actual conditions such as exposure, population shift, recovery, and susceptibility, the SI model has been expanded to SIR, SIER, Susceptible-Infected-Recovered-Dead (SIRD), and many more [348]. Because they are based on human understanding, these models are simple, explainable, and intuitive. However, they have limitations in describing the long-term dynamics of epidemic outbreaks, particularly when external influences substantially influence the dynamics.

The government's preventative and control actions, as well as public knowledge, all had a part in the spread of the pandemic. The availability of transparent pandemic reporting and preventive and control measures have hastened the virus's spread [312]. As a result, the pandemic historical data alone is not enough to provide reliable predictions. This chapter developed a data-driven epidemiological model for public health crises. We may overcome the limitations of classic epidemiological models that employ only a single component by integrating features from pandemic-related online information.

To address this issue, we use an LSTM network [184] with a PAN-LDA [208] module in our epidemiological model to revise the infection rate and increase the model's prediction accuracy. This article introduces the SIR-based model, incorporated with the PAN-LDA and the LSTM module for forecasting COVID-19.

7.2. Background

7.2.1. Susceptible-Infectious-Recovered Model

One of the most well-known epidemic models is the SIR model. In the SIR model, the population N is divided into three states: susceptible, $S(t)$, infectious, $I(t)$, or recovered, $R(t)$, at timestamp t. According to the model, the individual can go through two transitions, i.e., from S to I, and then from I to R. The parameters of the SIR model are the rate at which susceptible hosts become infected β and the rate at which infectious individuals recover r. The equations for the SIR model can then be established as:

$$\frac{dS(t)}{dt} = -\beta S(t) \tag{7.1}$$

$$\frac{dI(t)}{dt} = \beta S(t) - rI(t) \tag{7.2}$$

$$\frac{dR(t)}{dt} = rI(t) \tag{7.3}$$

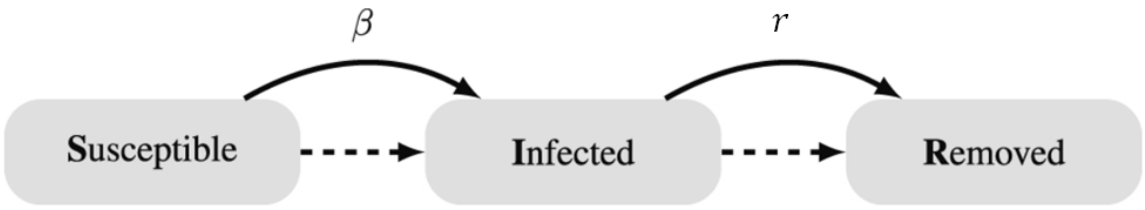


Figure 7.1 Compartmental Diagram for SIR Model

7.2.2. LSTM Network Model

Recurrent Neural Networks are well suited for sequence processing because they analyze the temporal behavior of a particular time series [176]. RNNs have a context layer that acts as memory, projecting information from the present state into future states, and then an output layer. Although there are other RNN designs, the Elman RNN [349] is one of the first and has been widely used to describe temporal sequences and dynamical systems. However, the fundamental limitation of RNNs is the vanishing gradient issue, which makes training challenging [286]. To address this issue, LSTM, a classic RNN, is frequently used [176], which is now widely employed in numerous disciplines such as text recognition, finance, and industrial engineering. In this chapter, LSTM is used as a model to optimize the COVID-19 confirmed cases predicted by the epidemic model.

Figure 3.2 depicts the construction of an LSTM block. Each LSTM block consists of a cell, an input gate, a forget gate, and an output gate. Each of the three gates performs a distinct function:

• The forget gate determines how much of the prior data will be forgotten and how much will be used in the following steps.

• The input gate determines what relevant information is fed into the cell state.

• The output gate determines the value of the next hidden state.

7.2.3. PAN-LDA Model

The government's prevention and control initiatives have had a major impact on the prevention of the epidemic's evolution, and media reporting of the pandemic, implementation of control and prevention measures, and reinforcement of citizens' prevention awareness have increased virus containment. To incorporate the impact of prevention and control measures by the government and public awareness, we collect pandemic-related data from various online platforms. In order to extract the relevant features from the collected textual data, PAN-LDA, an LDA-based feature extraction model, is employed. The PAN-LDA model takes the changes in historical time series into account along with textual data for performing feature selection in text and data mining, thereby increasing pandemic time series prediction performance.

7.3. Methodology

7.3.1. Framework of the Model

This paper proposed an epidemic model for COVID-19 prediction based on the SIR model. Furthermore, the LSTM method is utilized to optimize the epidemic model's infection rate and is paired with the SIR model to predict the number of infected patients. The model network diagram and the interaction among the individual components are shown in Figure 7.2.

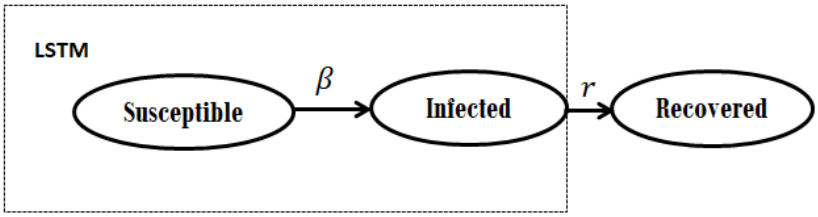


Figure 7.2 The Flow Diagram of the Proposed Epidemic Model

In addition, the LSTM model is used to revise the infection rate deviation calculated by the SIR model and is coupled with the SIR model to estimate the number of confirmed cases. This research employs the PAN-LDA model to extract features from relevant textual data to analyze the impact of preventive and control measures, transparent reporting on online platforms, and public awareness related to epidemic prevention. The collected features are then coupled with the LSTM method to update the infection rate deviation calculated by the epidemic model, which could forecast the number of infected cases. The framework of the epidemic prediction model is shown in Figure 7.3.

7.3.2. Prediction of the COVID-19 Cases

The traditional epidemiological models use historical data to forecast the infectious disease spread but do not take into account other variables such as media reporting and preventative and control efforts. As a result, known data-driven models must be incorporated to optimize the parameters of epidemic models. The LSTM network is a common method for modeling the hidden variables that are typically used for prediction, such as the number of potentially infected persons. However, experiments have demonstrated that the LSTM model alone cannot accurately estimate the number of infected patients. Given that preventative and control efforts and public knowledge of the disease play an important role in the transmission of infectious diseases, therefore, in this research, we used the PAN-LDA model to extract features from social media posts and news items related to the COVID-19 pandemic. The LSTM network is then embedded with these features, and the number of infected patients is then predicted by updating

the infection rate calculated by the classic SIR model. In order to increase the accuracy of epidemic prediction, the proposed approach updates the infection rate using news information and social media data.

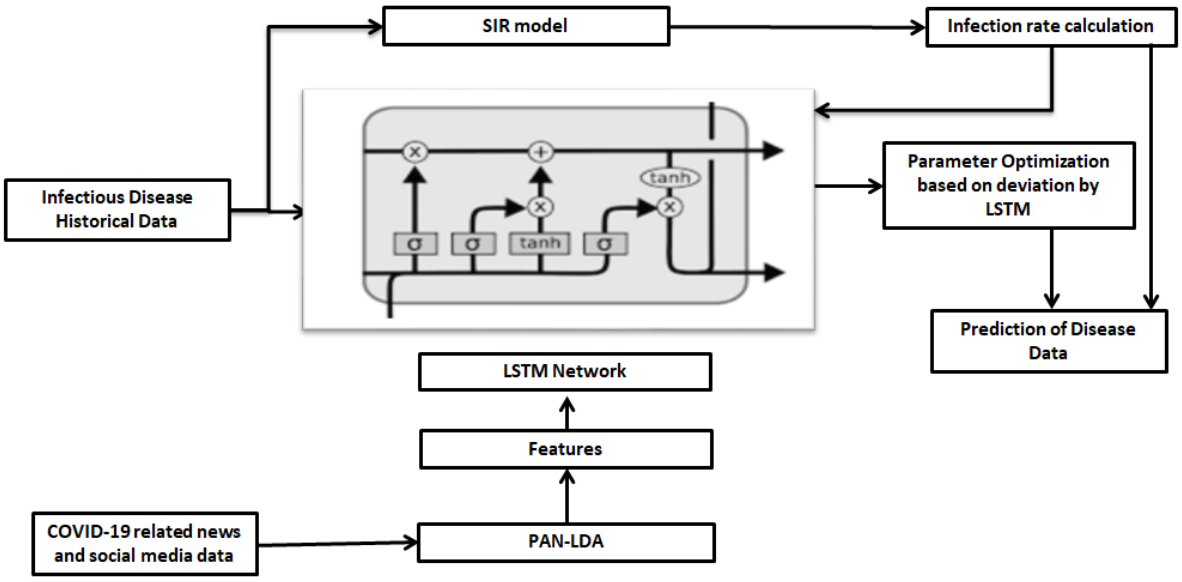


Figure 7.3 COVID-19 Prediction Model using Historical and Textual Data

7.3.3. Textual Feature Extraction

Twitter

Twitter has over 150 million active users globally. The platform enables researchers to mine the tweets that users submit online using publicly accessible Application Programming Interface (API) services, in accordance with the privacy constraints specified by the platform programmers. The Tweets are collected from the official Twitter account of the Ministry of Health & Family Welfare, Government of India [350] from March 1, 2021, to March 16, 2021. We then preprocessed the data by removing duplicate entries, filtering the irrelevant content, retaining only COVID-19-related content, and removing text in non-English languages, URLs, emojis, and punctuation.

Reddit

Reddit is a publicly available website. Reddit consists of distinct communities or subgroups that differ in the topic, called subreddits. Reddit users may produce original posts in a specific subreddit and comment on submissions. Using Python Reddit API Wrapper, we gathered coronavirus–related comments from 8

various subreddits, including r/indiacorona, r/CoronavirusIndia, r/covidIndia, r/COVID19, r/Coronavirus, r/COVID19support, r/nCoV, and r/CoronaVirus2019nCoV from March 1, 2021, to May 16, 2021.

Google News

Lastly, we collected English web news from Google search engine using the search terms "covid", "coronavirus", "corona", "India", and "news" to retrieve the relevant news articles. From this search, we collected news articles in the English language only.

Later the duplicates and irrelevant data that did not contain COVID-19-related information were filtered out from all three datasets. The data are classified by date, and the average of all features from the textual data of the day t is presented as a single textual feature vector for that day. The textual data is then preprocessed, and the bag-of-words representation of the textual data on day t is presented as input to the PAN-LDA, which is then converted into feature vectors.

7.3.4. LSTM Network Based on Textual Features and İnfection Rate

DNNs can fit complex distributions but tends to overfit without adequate observation. Infection rate-based epidemic models cannot foresee policy changes or emergencies and cannot make short-term predictions. As a result, we introduce the LSTM module based on textual features to mimic social media and existing policy, as illustrated in Figure 7.3. to assure short and long-term stability.

In the proposed model, we consider $\beta(t)$ as the real infection rate and that the regressed infection rate using an exponential function is $\hat{\beta}(t)$. The neural network is used to anticipate the difference between the actual and regress infection rates. To account for the influence of news and policy, we mix the textual features established in the preceding section with the infection rate bias. We utilized the LSTM network for encoding hidden states and temporal information. We use a single-layer perception model to convert infection rates and textual features into vectors.

Given infection features $s1$ and textual features $s2$, with $w1$ and $w2$ being the weights of the first two perceptron models. The convolution function g(.) followed by leaky ReLU is as follows:

$$f1 = g(s1; w1) \tag{7.4}$$

$$f2 = g(s2; w2) \tag{7.5}$$

The $f1$ and $f2$ are combined into a mixed feature f. Let $f(t)$ be the combined feature at timestamp t, considering $h(t-1)$ as the hidden state at timestamp $t-1$. The following $lstm()$ includes the fully connected layer and LSTM module that converts the hidden state into prediction. Then:

$$\big(x(t), h(t)\big) = lstm(f(t), h(t-1); wn) \tag{7.6}$$

Where $x(t)$ is the output, $h(t)$ is the new hidden state, and wn is the network's weight; the Adam optimizer is used as the optimization approach. The mean squared error between prediction and actual is used as the loss function.

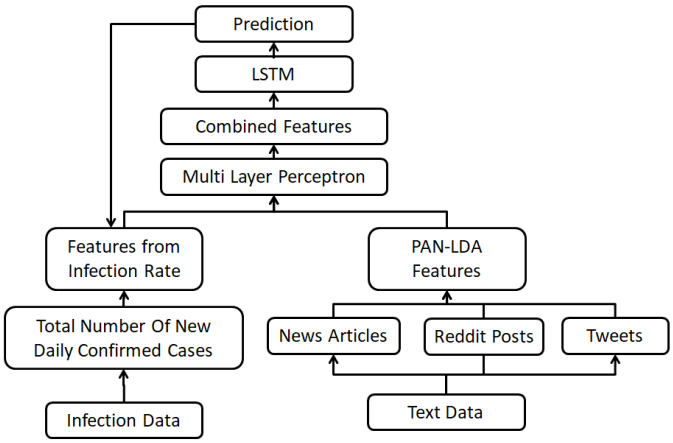


Figure 7.4 LSTM Network Based on Textual Features and Infection Rate

7.4. Results

This section validates the proposed model for four Indian states: Maharashtra, Kerala, Karnataka, and Delhi. The number of infected cases handled between March 1, 2021, to May 16, 2021, is utilized as training data to forecast the number of infections between May 17, 2021, and May 24, 2021.

We compared the classic SIR model, SIR+ LSTM, and the proposed model to validate the efficacy of the proposed model. We also utilize the proposed epidemic model to analyze the applicability of incorporating data from multiple social media platforms for pandemic prediction. In order to explore the effects of social media data, we compare the SIR+LSTM+features from Twitter and SIR+LSTM+features from Google news, Reddit posts, and Twitter. We compare the daily forecast, MAPE, and R^2 for Maharashtra, Kerala, Karnataka, and Delhi.

$$R^2 = \left(\frac{\sum_{i=1}^{N}(a_i - \bar{a})(f_i - \bar{f})}{\sqrt{\sum_{i=1}^{N}(a_i - \bar{a})^2}\sqrt{\sum_{i=1}^{N}(f_i - \bar{f})^2}} \right)^2 \tag{7.7}$$

$$MAPE = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{a_i - f_i}{a_i} \right| \tag{7.8}$$

where a_i represents the observed values and the f_i are the corresponding forecasted values on the i^{th} day respectively. $\bar{a}$ and $\bar{f}$ denotes the mean value of actual and forecasted values. N represents the total number of forecast days. Tables 7.1-7.4 indicate the outcomes of the comparison.

Table 7.1 Predicted and Actual Cases in Maharashtra

Date	SIR	SIR+LSTM	SIR+LSTM+TW	SIR+LSTM+ GN+RD+TW	Actuals
17 May	31691889	31276786	31477819	31981811	31338407
18 May	32976321	31868768	31478778	31989318	31588717
19 May	31246499	31978978	31898988	31865356	31874364
20 May	27086291	31676522	31898889	32899119	32154275
21 May	26908162	31786763	31190922	32991388	32441776
22 May	33798249	31786763	31909111	32889191	32723361
23 May	33036852	32878117	31998989	32777118	33013516
24 May	35209522	31878833	32099933	32617819	33277290
R^2	0.3250	0.5686	0.5718	**0.7113**	
MAPE	5.9039	1.5488	1.8274	**1.3205**	

Table 7.2 Predicted and Actual Cases in Kerala

Date	SIR	SIR+LSTM	SIR+LSTM+TW	SIR+LSTM+ GN+RD+TW	Actuals
17 May	18538374	17546464	18767676	18544535	18014842
18 May	18172927	17766567	18672376	17888774	18149395
19 May	18268519	17667656	18367637	18467643	18289940
20 May	18000224	17875564	18444988	18988665	18421465
21 May	18313600	18877654	18177378	18878765	18555023
22 May	18657837	17878667	18773677	18989898	18681051
23 May	19368790	17786657	18387322	19889811	18794256
24 May	19540265	18797878	19838787	18985676	18881587
R^2	0.7198	0.5878	0.3493	**0.7526**	
MAPE	1.6762	2.8641	2.1723	**2.2756**	

Table 7.3 Predicted and Actual Cases in Karnataka

Date	SIR	SIR+LSTM	SIR+LSTM+TW	SIR+LSTM+ GN+RD+TW	Actuals
17 May	27744501	27978176	27561655	27512454	27976933
18 May	28057856	27798789	27461767	27655656	28070180
19 May	27764393	27089181	28287722	27674445	28199718
20 May	28373754	27871117	28676655	28234356	28320429
21 May	28020899	27977181	28475544	28443546	28453442
22 May	28235857	28889718	28564546	28554344	28582203
23 May	28360876	27987917	28934118	28674677	28707320
24 May	28772799	28711881	28544454	28777555	28816043
R^2	0.8205	0.6154	0.8219	**0.9702**	
MAPE	0.8369	1.5141	0.8867	**0.7103**	

Table 7.4 Predicted and Actual Cases in Delhi

Date	SIR	SIR+LSTM	SIR+LSTM+TW	SIR+LSTM+ GN+RD+TW	Actuals
17 May	18294693	17876756	18467652	18254334	18342482
18 May	18412220	17861866	18586754	18223567	18407486
19 May	18478783	18678564	18345445	18635678	18474059
20 May	18409091	18643378	18476532	18434677	18532803
21 May	17922381	18088867	18434546	18543357	18595993
22 May	19007880	17866549	18544357	18585428	18659148
23 May	18694718	18897776	18765544	18654954	18727191
24 May	19163520	19877887	19554677	18533803	18788697
R^2	0.6309	0.6760	0.6830	**0.7219**	
MAPE	1.0798	2.6111	1.0526	**0.6630**	

Tables 7.1-7.4 show that the proposed model significantly outperforms the classic SIR model. Also, it can be noticed that the inclusion of COVID-19-related data from multiple sources has increased the prediction accuracy. The model that incorporates textual information from several online sources outperforms the other models in terms of precision. This discovery demonstrates that language features give extra information and guide disease prediction.

7.5. Chapter Summary

This chapter aims to predict the COVID-19 infected cases by proposing a SIR-based model with LSTM incorporating textual features from multiple social media platforms and historical time-series data to forecast the trend of COVID-19. In this study, the PAN-LDA model is used to evaluate and extract COVID-19-related news and awareness information from numerous online sources, which is then encoded into semantic characteristics. The features extracted are then incorporated into the LSTM model to revise the infection rate provided by the SIR model. The prediction results of the model are very consistent, demonstrating that the proposed model can forecast infection cases, and textual information processing of related news helps increase the model's accuracy. This research also shows that the openness, transparency, and efficiency of releasing data are essential for establishing a modern epidemic prevention system.

Chapter 8 CONCLUSION AND FUTURE SCOPE

This chapter presents a comprehensive overview of the numerous ways for public health surveillance, pandemic prediction using online content, and machine learning techniques. Section 8.1 provides an overview of the research contributions, while Section 8.2 discusses the limitations of the proposed methods. Section 8.3 delves into the future elements of the proposed work.

8.1. Research Summary

Our work introduces several new epidemic models using machine learning techniques for predicting outbreaks and public health surveillance. This thesis about public health surveillance presents my work on improving the prediction performance of epidemic models using online content and machine learning techniques. This thesis also includes my work on developing novel modeling methods and applying them to the challenging tasks of epidemic prediction and identification of risk factors. As per the literature review, various studies have performed public health surveillance but have some constraints and limited accuracy. The biggest challenge with public health surveillance is the unavailability of a specific and correct dataset for timely and accurate time-series prediction of the disease. Although various epidemic models have been proposed yet, to the best of our knowledge, the inclusion of online textual content in the proposed models was not considered for prediction purposes.

In order to achieve the RO1, i.e., to enhance the predictive performance of a public health event, we propose an epidemic compartmental model. For many years, compartmental models were used to predict infectious diseases efficiently and effectively. The proposed epidemic model is based on the SIR model, incorporating the "quarantined" and "vaccinated" compartments. The proposed model implements the evolutionary algorithm, LSTM neural network, for optimizing the model's parameters and resulting infection rate. The extracted features from disease-related news and social media postings are incorporated using the LDA-based PAN-LDA model to consider the impact of government control measures, transparent media coverage, and increased public knowledge about epidemic prevention. Chapter 3 goes over the whole description of the proposed model in detail.

For achieving RO2, first, we analyzed the correlation of the various non-medical health determinant factors such as socioeconomic, environmental, behavioral, and demographic factors with the spread of the pandemic. In our case, we explore the association of risk factors with the deadly coronavirus-2019 pandemic. We also explored the spatial and temporal differentiation of the pandemic spread, taking India as a study area. As the early stage of the infectious pandemic is associated with human mobility and environmental factors, we explored the impact of related multi-source data on the daily infected cases.

We also proposed a short-term fixed effect multiple regression model to predict confirmed cases in the early stages of the COVID-19 second wave. We also investigated the probability of the upcoming wave of the pandemic based on parameters associated with the pandemic transmission during the initial phases.

In order to attain RO3, two solutions have been proposed. In the first solution, we proposed a document representation method that creates concepts through clustering semantically and morphologically similar word vectors generated from word2vec and KPCA for document vector representation. With the appropriate weighting scheme, called modified-tfidf, the proposed approach provides a better document representation. The reduction of features dimensionality is also discussed using the autoencoders and self-training clustering mechanism. Another solution is based on the LDA-based model, called PAN-LDA, which offers better features extracted from online news articles and historical data to Machine Learning algorithms in order to improve time series prediction.

Since the number of individuals using social media to transmit information has increased dramatically, researchers have had a surge of interest in analyzing social media activity for public health purposes. Therefore, to achieve RO4, we conducted a literature review of surveillance systems in the health informatics sector using social media. Later, to show the applicability of using multiple social media platforms simultaneously, we presented a novel epidemic prediction model in which data from multiple online platforms was utilized to forecast infectious disease cases. The model is built on the SIR and LSTM models incorporating the features extracted from online content using the PAN-LDA model. Our experimental results show that incorporating features from multiple online platforms leads to a more accurate time-series prediction of the number of infected cases of the pandemic.

The contributions of this thesis are presented at the intersection of machine learning, online textual content, and public health events. The main contribution of this thesis is the machine learning models in which features from both historical data, and social media data are fed to markedly enhance the outbreak prediction. This, in turn, has the potential to improve public health decisions. This research shows that introduction of social media content about public health events along with historical data has improved the prediction performance of machine learning techniques. The research also indicates that efficiency, accessibility, and transparency in data release are critical for developing a contemporary epidemic prevention system. Overall, the proposed work is mainly suitable for tracking and predicting novel infectious diseases, which can provide significant recommendations to disease makers and epidemiologists. In this manner, they may establish proper policies to immediately and efficiently prevent and manage the outbreak and save as many lives as possible. In general, this work has aided the advancement of ML for public health applications.

8.2. Limitations of the Work

No one is perfect in the world, and every study has certain limits and constraints. This work is also subject to the following limitations:

- One of the most difficult challenges with public health surveillance is the unavailability of accurate data at the early stages of an outbreak. The underreporting of the infected cases may not fully reflect the true condition of a pandemic, thereby affecting the accuracy of the results.
- Another challenge of public health surveillance using online data is the reliability of the data. The data posted on the social media platforms also contains fake news, which affects the accuracy of the forecasting results. The size of the dataset also poses a challenge due to the unlimited amount of data that millions of members publish on a regular basis.
- The data for non-medical health determinants risk factors for some regions of the study area are not accessible.
- The assumption of a constant growth rate is a significant limitation, despite the fact that the infectious disease growth rate varies over time and can be influenced by a variety of epidemiological, socioeconomic, and health factors, resulting in a change in growth rate over time for different regions.

8.3. Future Aspects

Following are the future perspective of the work:

- A more in-depth analysis of choosing words from the vocabulary to produce the projection matrix required to generate KPCA embeddings.
- Considering the pandemic spreading on multi-layers networks and hybrid intelligent algorithms.
- Recent advances in the use of neural embedding and deep learning approaches may be used to provide enhanced feature representations or document clustering.
- Employing word and document embeddings in deep clustering and topic modeling algorithms as pre-trained initial layers.
- The work may be expanded to uncover data by utilizing user posts such as images, audio, and videos.
- Data from other social media platforms, such as Facebook, Instagram, YouTube, and others, may also be incorporated to provide more promising public health surveillance.
- Last but not the least, Fake news detection can be explored while dealing with social media data.

References

[1] Public Health Surveillance at CDC | CDC, (n.d.). https://www.cdc.gov/surveillance/improving-surveillance/Public-health-surveillance.html (accessed March 5, 2022).

[2] WHO, Surveillance In Emergencies, 2021. https://www.who.int/emergencies/surveillance (accessed January 4, 2022).

[3] S.L. Groseclose, D.L. Buckeridge, Public Health Surveillance Systems: Recent Advances in Their Use and Evaluation, Annu. Rev. Public Health. 38 (2017) 57–79. https://doi.org/10.1146/annurev-publhealth-031816-044348.

[4] B.C.K. Choi, A.W.P. Pak, Lessons for surveillance in the 21st century: A historical perspective from the past five millennia, Soz. Praventivmed. 46 (2001) 361–368. https://doi.org/10.1007/BF01321662.

[5] G. Marks, W.K. Beatty, Epidemics, (1976) 323.

[6] B.C.K. Choi, The Past, Present, and Future of Public Health Surveillance, Scientifica (Cairo). 2012 (2012) 1–26. https://doi.org/10.6064/2012/875253.

[7] Eylenbosch: Historical aspects - Google Scholar, (n.d.). https://scholar.google.com/scholar_lookup?title=Historical aspects&author=W. J. Eylenbosch &author=N. D. Noah&publication_year=1988 (accessed March 5, 2022).

[8] U.O. Department Health, H. Services, C. for Disease Control, Principles of Epidemiology in Public Health Practice, Third Edition: An Introduction, (2006).

[9] P. Nsubuga, M.E. White, S.B. Thacker, M.A. Anderson, S.B. Blount, C. V. Broome, T.M. Chiller, V. Espitia, R. Imtiaz, D. Sosin, D.F. Stroup, R. V. Tauxe, M. Vijayaraghavan, M. Trostle, Chapter 53. Public Health Surveillance: A Tool for Targeting and Monitoring Interventions, in: D.T. Jamison, J.G. Breman, A.R. Measham, G. Alleyne, M. Claeson, D.B. Evans, P. Jha, A. Mills, P. Musgrove (Eds.), Dis. Control Priorities Dev. Ctries. (2nd Ed., The International Bank for Reconstruction and Development / The World Bank, 2006: pp. 997–1016. https://doi.org/10.1596/978-0-8213-6179-5/chpt-53.

[10] W.H. Foege, R.C. Hogan, L.H. Newton, Surveillance projects for selected diseases, Int. J. Epidemiol. 5 (1976) 29–37. https://doi.org/10.1093/ije/5.1.29.

[11] A. Garcia-abreu, W. Halperin, I. Danel, Public Health Surveillance Toolkit. A guid for busy task managers, Wold Bank. (2002).

[12] R. Gilbert, S.J. Cliffe, Public health surveillance, in: Public Heal. Intell. Issues Meas. Method, Springer, Cham, 2016: pp. 91–110. https://doi.org/10.1007/978-3-319-28326-5_5.

[13] S.B. Thacker, D.F. Stroup, Future directions for comprehensive public health surveillance and health information systems in the united states, Am. J. Epidemiol. 140 (1994) 383–397. https://doi.org/10.1093/oxfordjournals.aje.a117261.

[14] L. Lenert, D.N. Sundwall, Public health surveillance and meaningful use regulations: A crisis of opportunity, Am. J. Public Health. 102 (2012) 1–7. https://doi.org/10.2105/AJPH.2011.300542.

[15] T.D. Vacalis, C. Bartlett, C.G. Shapiro, Electronic communication and the future of international public health surveillance., Emerg. Infect. Dis. 1 (1995) 34–35. https://doi.org/10.3201/eid0101.950108.

[16] Centers for Disease Control (CDC), National Electronic Telecommunications System for Surveillance--United States, 1990-1991., Morb. Mortal. Wkly. Rep. 40 (1991) 502–503. https://sci-hub.do/https://pubmed.ncbi.nlm.nih.gov/1649378/ (accessed March 5, 2022).

[17] R.R. German, L.M. Lee, J.M. Horan, R.L. Milstein, C.A. Pertowski, M.N. Waller, Guidelines Working Group Centers for Disease Control and Prevention (CDC), Updated guidelines for evaluating public health surveillance systems: recommendations from the Guidelines Working Group., MMWR. Recomm. Reports Morb. Mortal. Wkly. Report. Recomm. Reports. 50 (2001) 1–35; quiz CE1-7. https://sci-hub.do/https://stacks.cdc.gov/view/cdc/7555 (accessed March 5, 2022).

[18] V. Della Mea, What is e-health (2): The death of telemedicine?, J. Med. Internet Res. 3 (2001) 6–7. https://doi.org/10.2196/jmir.3.2.e22.

[19] L. Samaras, M.A. Sicilia, E. García-Barriocanal, Predicting epidemics using search engine data: a comparative study on measles in the largest countries of Europe, BMC Public Health. 21 (2021) 1–14. https://doi.org/10.1186/s12889-020-10106-8.

[20] S. Valentin, A. Mercier, R. Lancelot, M. Roche, E. Arsevska, Monitoring online media reports for early detection of unknown diseases: Insight from a retrospective study of COVID-19 emergence, Transbound. Emerg. Dis. 68 (2021) 981–986. https://doi.org/10.1111/tbed.13738.

[21] S.M. van Rooden, O. Aspevall, E. Carrara, S. Gubbels, A. Johansson, J.C. Lucet, S. Mookerjee, Z.R. Palacios-Baena, E. Presterl, E. Tacconelli, M. Abbas, M. Behnke, P. Gastmeier, M.S.M. van Mourik, Governance aspects of large-scale implementation of automated surveillance of healthcare-associated infections, Clin. Microbiol. Infect. 27 (2021) S20–S28. https://doi.org/10.1016/j.cmi.2021.02.026.

[22] G. Eysenbach, Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet, J. Med. Internet Res. 11 (2009) e1157. https://doi.org/10.2196/jmir.1157.

[23] D.B. Neill, New directions in artificial intelligence for public health surveillance, IEEE Intell.

Syst. 27 (2012) 56–59. https://doi.org/10.1109/MIS.2012.18.

[24] S.J. Yan, A.A. Chughtai, C.R. Macintyre, Utility and potential of rapid epidemic intelligence from internet-based sources, Int. J. Infect. Dis. 63 (2017) 77–87. https://doi.org/10.1016/j.ijid.2017.07.020.

[25] S.M. Noar, E. Leas, B.M. Althouse, M. Dredze, D. Kelley, J.W. Ayers, Can a selfie promote public engagement with skin cancer ?, Prev. Med. (Baltim). (2017) 1–4. https://doi.org/10.1016/j.ypmed.2017.10.038.

[26] L. Mollema, I.A. Harmsen, E. Broekhuizen, R. Clijnk, H. De Melker, T. Paulussen, G. Kok, R. Ruiter, E. Das, Disease Detection or Public Opinion Reflection ? Content Analysis of Tweets , Other Social Media , and Online Newspapers During the Measles Outbreak in the Netherlands in 2013, 17 (2013) 1–12. https://doi.org/10.2196/jmir.3863.

[27] I.C.H. Fung, C.H. Duke, K.C. Finch, K.R. Snook, P.L. Tseng, A.C. Hernandez, M. Gambhir, K.W. Fu, Z.T.H. Tse, Ebola virus disease and social media: A systematic review, Am. J. Infect. Control. 44 (2016) 1660–1671. https://doi.org/10.1016/j.ajic.2016.05.011.

[28] M. Conway, D. O'Connor, Social media, big data, and mental health: Current advances and ethical implications, Curr. Opin. Psychol. 9 (2016) 77–82. https://doi.org/10.1016/j.copsyc.2016.01.004.

[29] L. Fernandez-luque, M. Imran, Humanitarian Health Computing using Artificial Intelligence and Social Media: A Narrative Literature Review, Int. J. Med. Inform. (2018). https://doi.org/10.1016/j.ijmedinf.2018.01.015.

[30] J. O'Shea, Digital disease detection: A systematic review of event-based internet biosurveillance systems, Int. J. Med. Inform. 101 (2017) 15–22. https://doi.org/10.1016/j.ijmedinf.2017.01.019.

[31] H.A. Park, H. Jung, J. On, S.K. Park, H. Kang, Digital epidemiology: Use of digital data collected for non-epidemiological purposes in epidemiological studies, Healthc. Inform. Res. (2018). https://doi.org/10.4258/hir.2018.24.4.253.

[32] W.D. Jenkins, B. Wold, Use of the Internet for the surveillance and prevention of sexually transmitted diseases, Microbes Infect. 14 (2012) 427–437. https://doi.org/10.1016/j.micinf.2011.12.006.

[33] G.D. Haddow, K.S. Haddow, G.D. Haddow, K.S. Haddow, Chapter Eleven – Communicating During a Public Health Crisis, Disaster Commun. a Chang. Media World. (2014) 195–209. https://doi.org/10.1016/B978-0-12-407868-0.00011-2.

[34] E. Yom-tov, Ebola data from the Internet : An Opportunity for Syndromic Surveillance or a News Event ? Categories and Subject Descriptors, (n.d.) 115–119.

[35] J. Mowery, Twitter Influenza Surveillance: Quantifying Seasonal Misdiagnosis Patterns, Online J. Public Health Inform. 8 (2016). https://doi.org/10.5210/ojphi.v8i3.7011.

[36] T. Nguyen, M.E. Larsen, B.O. Dea, D.T. Nguyen, J. Yearwood, D. Phung, S. Venkatesh, H. Christensen, Kernel-based features for predicting population health indices from geocoded social media data, (2017). https://doi.org/10.1016/j.dss.2017.06.010.

[37] P. Kostkova, A Roadmap to Integrated Digital Public Health Surveillance : the Vision and the Challenges, (2013) 687–693.

[38] S. Chaudhary, S. Naaz, Use of Big Data in Computational Epidemiology for Public Health Surveillance, (2017) 150–155.

[39] H. Kwak, C. Lee, H. Park, S. Moon, What is Twitter, a Social Network or a News Media?, Arch. Zootec. 60 (2011) 297–300. https://doi.org/10.4321/S0004-05922011000200015.

[40] A. Stefanidis, E. Vraga, G. Lamprianidis, J. Radzikowski, P.L. Delamater, K.H. Jacobsen, D. Pfoser, A. Croitoru, A. Crooks, Zika in Twitter: Temporal variations of locations, actors, and concepts, JMIR Public Heal. Surveill. 3 (2017). https://doi.org/10.2196/publichealth.6925.

[41] K. Rudra, A. Sharma, N. Ganguly, M. Imran, Classifying Information from Microblogs during Epidemics, Proc. 2017 Int. Conf. Digit. Heal. - DH '17. (2017) 104–108. https://doi.org/10.1145/3079452.3079491.

[42] J.C. Bosley, N.W. Zhao, S. Hill, F.S. Shofer, D.A. Asch, L.B. Becker, R.M. Merchant, Decoding twitter: Surveillance and trends for cardiac arrest and resuscitation communication, Resuscitation. 84 (2013) 206–212. https://doi.org/10.1016/j.resuscitation.2012.10.017.

[43] K. Lee, A. Agrawal, A. Choudhary, Forecasting Influenza Levels Using Real-Time Social Media Streams, in: Proc. - 2017 IEEE Int. Conf. Healthc. Informatics, ICHI 2017, 2017: pp. 409–414. https://doi.org/10.1109/ICHI.2017.68.

[44] K. Lee, A. Agrawal, A. Choudhary, Mining social media streams to improve public health allergy surveillance, in: Proc. 2015 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2015, ACM, 2015: pp. 815–822. https://doi.org/10.1145/2808797.2808896.

[45] M.O. Edd, S.Y. Rn, What can we learn about the Ebola outbreak from tweets ?, Am. J. Infect. Control. 43 (2015) 563–571. https://doi.org/10.1016/j.ajic.2015.02.023.

[46] X. Dai, M. Bikdash, B. Meyer, From social media to public health surveillance: Word embedding based clustering method for twitter classification, in: Conf. Proc. - IEEE SOUTHEASTCON, IEEE, 2017. https://doi.org/10.1109/SECON.2017.7925400.

[47] N. El-Bathy, C. Gloster, G. Azar, M. El-Bathy, G. Stein, R. Stevenson, Intelligent surveillance lifecycle architecture for epidemiological data clustering using Twitter and novel genetic algorithm, in: IEEE Int. Conf. Electro Inf. Technol., IEEE, 2014: pp. 149–155. https://doi.org/10.1109/EIT.2014.6871753.

[48] Y. Khan, G.J. Leung, P. Belanger, E. Gournis, D.L. Buckeridge, L. Liu, Y. Li, I.L. Johnson,

Comparing Twitter data to routine data sources in public health surveillance for the 2015 Pan/Parapan American Games: an ecological study, Can. J. Public Heal. 109 (2018) 419–426. https://doi.org/10.17269/s41997-018-0059-0.

[49] K. Nargund, S. Natarajan, Public health allergy surveillance using micro-blogs, 2016 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2016. (2016) 1429–1433. https://doi.org/10.1109/ICACCI.2016.7732248.

[50] L. Sousa, R. de Mello, D. Cedrim, A. Garcia, P. Missier, A. Uchôa, A. Oliveira, A. Romanovsky, VazaDengue: An information system for preventing and combating mosquito-borne diseases with social networks, Inf. Syst. 75 (2018) 26–42. https://doi.org/10.1016/j.is.2018.02.003.

[51] K. Espina, M.R.J.E. Estuar, Infodemiology for Syndromic Surveillance of Dengue and Typhoid Fever in the Philippines, in: Procedia Comput. Sci., Elsevier B.V., 2017: pp. 554–561. https://doi.org/10.1016/j.procs.2017.11.073.

[52] A. Veloso, F. Ferraz, Dengue surveillance based on a computational model of spatio-temporal locality of Twitter, (2011).

[53] K. Talvis, K. Chorianopoulos, K.L. Kermanidis, Real-time monitoring of flu epidemics through linguistic and statistical analysis of twitter messages, Proc. - 9th Int. Work. Semant. Soc. Media Adapt. Pers. SMAP 2014. (2014) 83–87. https://doi.org/10.1109/SMAP.2014.38.

[54] V. Lampos, N. Cristianini, Tracking the flu pandemic by monitoring the social web, 2010 2nd Int. Work. Cogn. Inf. Process. CIP2010. (2010) 411–416. https://doi.org/10.1109/CIP.2010.5604088.

[55] A.A. Aslam, M.H. Tsou, B.H. Spitzberg, L. An, J.M. Gawron, D.K. Gupta, K.M. Peddecord, A.C. Nagel, C. Allen, J.A. Yang, S. Lindsay, The reliability of tweets as a supplementary method of seasonal influenza surveillance, J. Med. Internet Res. 16 (2014). https://doi.org/10.2196/jmir.3532.

[56] P. Kostkova, M. Szomszor, C. St. Luis, #Swineflu: The Use of Twitter as an EarlyWarning and Risk Communication, ACM Trans. Manag. Inf. Syst. 5 (2014) 1–25. https://doi.org/10.1145/2597892.

[57] A. McNeill, P.R. Harris, P. Briggs, Twitter Influence on UK Vaccination and Antiviral Uptake during the 2009 H1N1 Pandemic, Front. Public Heal. 4 (2016) 26. https://doi.org/10.3389/fpubh.2016.00026.

[58] A. Alessa, M. Faezipour, Preliminary Flu Outbreak Prediction Using Twitter Posts Classification and Linear Regression With Historical Centers for Disease Control and Prevention Reports: Prediction Framework Study, JMIR Public Heal. Surveill 2019;5(2)E12383 Https//Publichealth.Jmir.Org/2019/2/E12383. 5 (2019) e12383. https://doi.org/10.2196/12383.

[59] M.D. Shoaei, M. Dastani, The role of twitter during the COVID-19 crisis: A systematic literature review, Acta Inform. Pragensia. 9 (2020) 154–169. https://doi.org/10.18267/J.AIP.138.

[60] R.J. Medford, S.N. Saleh, A. Sumarsono, T.M. Perl, C.U. Lehmann, An “Infodemic”: Leveraging high-volume twitter data to understand early public sentiment for the Coronavirus disease 2019 outbreak, Open Forum Infect. Dis. 7 (2020). https://doi.org/10.1093/ofid/ofaa258.

[61] Y. Lama, T. Chen, M. Dredze, A. Jamison, S.C. Quinn, D.A. Broniatowski, Discordance between human papillomavirus twitter images and disparities in human papillomavirus risk and disease in the United States: Mixed-methods analysis, J. Med. Internet Res. 20 (2018) e10244. https://doi.org/10.2196/10244.

[62] N.J.L. Brown, J.C. Coyne, Does Twitter language reliably predict heart disease? A commentary on Eichstaedt et al. (2015a), PeerJ. 2018 (2018) e5656. https://doi.org/10.7717/peerj.5656.

[63] J.P.D. Guidry, Y. Jin, C.A. Orr, M. Messner, S. Meganck, Ebola on Instagram and Twitter: How health organizations address the health crisis in their social media engagement, Public Relat. Rev. 43 (2017) 477–486. https://doi.org/10.1016/j.pubrev.2017.04.009.

[64] E.K. Seltzer, E. Horst-Martz, M. Lu, R.M. Merchant, Public sentiment and discourse about Zika virus on Instagram, Public Health. 150 (2017) 170–175. https://doi.org/10.1016/j.puhe.2017.07.015.

[65] E.E. Arolas, F.G. Ladrón-de-Guevara, Towards an integrating crowdsourcing definition, 32 (2016) 189–200. https://doi.org/10.1177/016555150000000.

[66] N. EO., K. SA., B. JS., Online reports of foodborne illness capture foods implicated in official foodborne outbreak reports., Prev. Med. (Baltim). 67 (2014) 264–269. https://doi.org/10.1016/j.ypmed.2014.08.003.

[67] M.O. Lwin, S. Vijaykumar, O.N.N. Fernando, S.A. Cheong, V.S. Rathnayake, G. Lim, Y.L. Theng, S. Chaudhuri, S. Foo, A 21st century approach to tackling dengue: Crowdsourced surveillance, predictive mapping and tailored communication, Acta Trop. 130 (2014) 100–107. https://doi.org/10.1016/j.actatropica.2013.09.021.

[68] A. Ghenai, Y. Mejova, Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter, in: Proc. - 2017 IEEE Int. Conf. Healthc. Informatics, ICHI 2017, 2017: p. 518. https://doi.org/10.1109/ICHI.2017.58.

[69] P. Quade, E.O. Nsoesie, A platform for crowdsourced foodborne illness surveillance: Description of users and reports, JMIR Public Heal. Surveill. 3 (2017). https://doi.org/10.2196/publichealth.7076.

[70] N. Yang, X. Cui, C. Hu, W. Zhu, C. Yang, Chinese social media analysis for disease surveillance, in: Proc. - 2014 Int. Conf. Identification, Inf. Knowl. Internet Things, IIKI 2014, IEEE, 2014: pp. 17–21. https://doi.org/10.1109/IIKI.2014.11.

[71] Z. Ertem, D. Raymond, L.A. Meyers, Optimal multi-source forecasting of seasonal influenza,

PLoS Comput. Biol. 14 (2018) 1–16. https://doi.org/10.1371/journal.pcbi.1006236.

[72] K. Liu, L. Li, T. Jiang, B. Chen, Z. Jiang, Z. Wang, Y. Chen, J. Jiang, H. Gu, Chinese public attention to the outbreak of ebola in west africa: Evidence from the online big data platform, Int. J. Environ. Res. Public Health. 13 (2016). https://doi.org/10.3390/ijerph13080780.

[73] I.C.H. Fung, K.W. Fu, Y. Ying, B. Schaible, Y. Hao, C.H. Chan, Z.T.H. Tse, Chinese social media reaction to the MERS-CoV and avian influenza A(H7N9) outbreaks, Infect. Dis. Poverty. 2 (2013) 1–12. https://doi.org/10.1186/2049-9957-2-31.

[74] I.C.H. Fung, K.W. Fu, C.H. Chan, B.S.B. Chan, C.N. Cheung, T. Abraham, Z.T.H. Tse, Social media's initial reaction to information and misinformation on ebola, august 2014: Facts and rumors, Public Health Rep. 131 (2016) 461–473. https://doi.org/10.1177/003335491613100312.

[75] B. Chen, J. Shao, K. Liu, G. Cai, Z. Jiang, Y. Huang, H. Gu, J. Jiang, Does eating chicken feet with pickled peppers cause avian influenza? Observational case study on Chinese social media during the avian influenza a (h7n9) outbreak, JMIR Public Heal. Surveill. 4 (2018). https://doi.org/10.2196/publichealth.8198.

[76] H.A. Carneiro, E. Mylonakis, Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks, Clin. Infect. Dis. 49 (2009) 1557–1564. https://doi.org/10.1086/630200.

[77] J.D. Sharpe, R.S. Hopkins, R.L. Cook, C.W. Striley, Evaluating Google, Twitter, and Wikipedia as Tools for Influenza Surveillance Using Bayesian Change Point Analysis: A Comparative Analysis, JMIR Public Heal. Surveill. 2 (2016) e161. https://doi.org/10.2196/publichealth.5901.

[78] S. Ram, W. Zhang, M. Williams, Y. Pengetnze, Predicting asthma-related emergency department visits using big data, IEEE J. Biomed. Heal. Informatics. 19 (2015) 1216–1223. https://doi.org/10.1109/JBHI.2015.2404829.

[79] X. Zhou, J. Ye, Y. Feng, Tuberculosis surveillance by analyzing google trends, IEEE Trans. Biomed. Eng. 58 (2011) 2247–2254. https://doi.org/10.1109/TBME.2011.2132132.

[80] T.J. Bruno, K.H. Wertz, Prediction of influenza outbreaks by integrating Wikipedia article access logs and Google flu trend data, J. Chromatogr. A. 736 (1996) 175–184. https://doi.org/10.1109/BIBE.2015.7367640.

[81] H. Xue, Y. Bai, H. Hu, H. Liang, Influenza Activity Surveillance Based on Multiple Regression Model and Artificial Neural Network, IEEE Access. 6 (2017) 563–575. https://doi.org/10.1109/ACCESS.2017.2771798.

[82] D.W. Seo, S.Y. Shin, Methods using social media and search queries to predict infectious disease outbreaks, Healthc. Inform. Res. 23 (2017) 343–348. https://doi.org/10.4258/hir.2017.23.4.343.

[83] Y. Luo, D. Zeng, Z. Cao, X. Zheng, Y. Wang, Q. Wang, H. Zhao, Using multi-source web data for epidemic surveillance: A case study of the 2009 Influenza A (H1N1) pandemic in Beijing, Proc.

2010 IEEE Int. Conf. Serv. Oper. Logist. Informatics, SOLI 2010. (2010) 76–81. https://doi.org/10.1109/SOLI.2010.5551614.

[84] C.D. Corley, D.J. Cook, A.R. Mikler, K.P. Singh, Using web and social media for influenza surveillance, in: Adv. Exp. Med. Biol., 2010: pp. 559–564. https://doi.org/10.1007/978-1-4419-5913-3_61.

[85] Y.A. Strekalova, Emergent health risks and audience information engagement on social media, Am. J. Infect. Control. 44 (2016) 363–365. https://doi.org/10.1016/j.ajic.2015.09.024.

[86] S. Gittelman, V. Lange, C.A. Gotway Crawford, C.A. Okoro, E. Lieb, S.S. Dhingra, E. Trimarchi, A new source of data for public health surveillance: Facebook likes, J. Med. Internet Res. 17 (2015) e98. https://doi.org/10.2196/jmir.3970.

[87] C.H. Basch, C.E. Basch, K. V. Ruggles, R. Hammond, Coverage of the Ebola Virus Disease Epidemic on YouTube, Disaster Med. Public Health Prep. 9 (2015) 531–535. https://doi.org/10.1017/dmp.2015.77.

[88] A. Nerghes, P. Kerkhof, I. Hellsten, Early Public Responses to the Zika-Virus on YouTube : Prevalence of and Differences Between Conspiracy Theory and Informational Videos, 10th ACM Conf. OnWeb Sci. (2018) 127–134. https://doi.org/10.1145/3201064.3201086.

[89] S. Choi, J. Lee, M.G. Kang, H. Min, Y.S. Chang, S. Yoon, Large-scale machine learning of media outlets for understanding public reactions to nation-wide viral infection outbreaks, Methods. 129 (2017) 50–59. https://doi.org/10.1016/j.ymeth.2017.07.027.

[90] S.D. Young, N. Mercer, R.E. Weiss, E.A. Torrone, S.O. Aral, Using social media as a tool to predict syphilis, Prev. Med. (Baltim). 109 (2018) 58–61. https://doi.org/10.1016/j.ypmed.2017.12.016.

[91] C. Chew, G. Eysenbach, Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak, PLoS One. 5 (2010) 1–13. https://doi.org/10.1371/journal.pone.0014118.

[92] L. Chen, K.S.M.T. Hossain, P. Butler, N. Ramakrishnan, B.A. Prakash, Flu Gone Viral : Syndromic Surveillance of Flu on Twitter using Temporal Topic Models, (2014) 755–760. https://doi.org/10.1109/ICDM.2014.137.

[93] K. Byrd, A. Mansurov, O. Baysal, Mining Twitter data for influenza detection and surveillance, Proc. Int. Work. Softw. Eng. Healthc. Syst. - SEHS ’16. (2016) 43–49. https://doi.org/10.1145/2897683.2897693.

[94] Q. Zhang, K. Sun, M. Chinazzi, A.P.Y. Piontti, N.E. Dean, Di.P. Rojas, S. Merler, Di. Mistry, P. Poletti, L. Rossi, M. Bray, M.E. Halloran, I.M. Longini, A. Vespignani, Spread of Zika virus in the Americas, Proc. Natl. Acad. Sci. U. S. A. 114 (2017) E4334–E4343. https://doi.org/10.1073/pnas.1620161114.

[95] O.B. Da'ar, F. Yunus, N. Md. Hossain, M. Househ, Impact of Twitter intensity, time, and location on message lapse of bluebird's pursuit of fleas in Madagascar, J. Infect. Public Health. 10 (2017) 396–402. https://doi.org/10.1016/j.jiph.2016.06.011.

[96] N. Thapen, D. Simmie, C. Hankin, J. Gillard, DEFENDER: Detecting and forecasting epidemics using novel data-analytics for enhanced response, PLoS One. 11 (2016) 1–19. https://doi.org/10.1371/journal.pone.0155417.

[97] T. Tran, K. Lee, Understanding citizen reactions and Ebola-related information propagation on social media, in: Proc. 2016 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2016, IEEE, 2016: pp. 106–111. https://doi.org/10.1109/ASONAM.2016.7752221.

[98] J. Du, L. Tang, Y. Xiang, D. Zhi, J. Xu, H.Y. Song, C. Tao, Public perception analysis of tweets during the 2015 measles outbreak: Comparative study using convolutional neural network models, J. Med. Internet Res. 20 (2018) 1–11. https://doi.org/10.2196/jmir.9413.

[99] K.W. Fu, H. Liang, N. Saroha, Z.T.H. Tse, P. Ip, I.C.H. Fung, How people react to Zika virus outbreaks on Twitter? A computational content analysis, Am. J. Infect. Control. 44 (2016) 1700–1702. https://doi.org/10.1016/j.ajic.2016.04.253.

[100] R. Gaspar, S. Gorjão, B. Seibt, L. Lima, J. Barnett, A. Moss, J. Wills, Tweeting during food crises: A psychosocial analysis of threat coping expressions in Spain, during the 2011 European EHEC outbreak, Int. J. Hum. Comput. Stud. 72 (2014) 239–254. https://doi.org/10.1016/j.ijhcs.2013.10.001.

[101] L. Tang, B. Bie, D. Zhi, Tweeting about measles during stages of an outbreak: A semantic network approach to the framing of an emerging infectious disease, Am. J. Infect. Control. 46 (2018) 1375–1380. https://doi.org/10.1016/j.ajic.2018.05.019.

[102] X. Gui, Y. Kou, K.H. Pine, Y. Chen, Managing uncertainty: Using social media for risk assessment during a public health crisis, in: Conf. Hum. Factors Comput. Syst. - Proc., 2017: pp. 4520–4533. https://doi.org/10.1145/3025453.3025891.

[103] I.C.H. Fung, K.W. Fu, C.H. Chan, B.S.B. Chan, C.N. Cheung, T. Abraham, Z.T.H. Tse, Social media's initial reaction to information and misinformation on ebola, august 2014: Facts and rumors, Public Health Rep. 131 (2016) 461–473. https://doi.org/10.1177/003335491613100312.

[104] L. Tang, B. Bie, S.E. Park, D. Zhi, Social media and outbreaks of emerging infectious diseases: A systematic review of literature, Am. J. Infect. Control. 46 (2018) 962–972. https://doi.org/10.1016/j.ajic.2018.02.010.

[105] E. Hagg, V.S. Dahinten, L.M. Currie, The emerging use of social media for health-related purposes in low and middle-income countries: A scoping review, Int. J. Med. Inform. 115 (2018) 92–105. https://doi.org/10.1016/j.ijmedinf.2018.04.010.

[106] C. Robertson, L. Yee, Avian influenza risk surveillance in North America with online media, PLoS One. 11 (2016). https://doi.org/10.1371/journal.pone.0165688.

[107] R.A. Calix, R. Gupta, M. Gupta, K. Jiang, Deep gramulator: Improving precision in the classification of personal health-experience tweets with deep learning, Proc. - 2017 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2017. 2017-Janua (2017) 1154–1159. https://doi.org/10.1109/BIBM.2017.8217820.

[108] D. Godfrey, C. Johns, C. Meyer, S. Race, C. Sadek, A Case Study in Text Mining: Interpreting Twitter Data From World Cup Tweets, (2014). http://arxiv.org/abs/1408.5427 (accessed September 8, 2020).

[109] G. Blouin-Genest, A. Miller, The politics of participatory epidemiology: Technologies, social media and influenza surveillance in the US, Heal. Policy Technol. 6 (2017) 192–197. https://doi.org/10.1016/j.hlpt.2017.02.001.

[110] T. Bodnar, M. Salathé, Validating models for disease detection using twitter, in: WWW 2013 Companion - Proc. 22nd Int. Conf. World Wide Web, 2013: pp. 699–702. https://doi.org/10.1145/2487788.2488027.

[111] A.A. Bharambe, D.R. Kalbande, Techniques and Approaches for Disease Outbreak Prediction, (2016) 100–102. https://doi.org/10.1145/2909067.2909085.

[112] J.R. Cataldi, A.F. Dempsey, S.T. O'Leary, Measles, the media, and MMR: Impact of the 2014–15 measles outbreak, Vaccine. 34 (2016) 6375–6380. https://doi.org/10.1016/j.vaccine.2016.10.048.

[113] Y. Kou, X. Gui, Y. Chen, K. Pine, Conspiracy Talk on Social Media: Collective Sensemaking during a Public Health Crisis, Proc. ACM Human-Computer Interact. 1 (2017) 1–21. https://doi.org/10.1145/3134696.

[114] E.O. Nsoesie, L. Flor, J. Hawkins, A. Maharana, T. Skotnes, F. Marinho, J.S. Brownstein, Social Media as a Sentinel for Disease Surveillance: What Does Sociodemographic Status Have to Do with It?, PLoS Curr. (2016). https://doi.org/10.1371/currents.outbreaks.cc09a42586e16dc7dd62813b7ee5d6b6.

[115] L.E. Charles-Smith, T.L. Reynolds, M.A. Cameron, M. Conway, E.H.Y. Lau, J.M. Olsen, J.A. Pavlin, M. Shigematsu, L.C. Streichert, K.J. Suda, C.D. Corley, Using social media for actionable disease surveillance and outbreak management: A systematic literature review, PLoS One. 10 (2015). https://doi.org/10.1371/journal.pone.0139701.

[116] G. Barata, K. Shores, J.P. Alperin, Local chatter or international buzz? Language differences on posts about Zika research on Twitter and Facebook, PLoS One. 13 (2018). https://doi.org/10.1371/journal.pone.0190482.

[117] M.U. Ilyas, J.S. Alowibdi, Disease Tracking in GCC Region Using Arabic Language Tweets, in:

Web Conf. 2018 - Companion World Wide Web Conf. WWW 2018, Association for Computing Machinery, Inc, 2018: pp. 417–423. https://doi.org/10.1145/3184558.3186357.

[118] I.C.H. Fung, J. Zeng, C.H. Chan, H. Liang, J. Yin, Z. Liu, Z.T.H. Tse, K.W. Fu, Twitter and Middle East respiratory syndrome, South Korea, 2015: A multi-lingual study, Infect. Dis. Heal. 23 (2018) 10–16. https://doi.org/10.1016/j.idh.2017.08.005.

[119] R. McKee, Ethical issues in using social media for health and health care research, Health Policy (New. York). 110 (2013) 298–301. https://doi.org/10.1016/j.healthpol.2013.02.006.

[120] E.M. Eggleston, E.R. Weitzman, Innovative uses of electronic health records and social media for public health surveillance, Curr. Diab. Rep. 14 (2014). https://doi.org/10.1007/s11892-013-0468-7.

[121] M.A. Mayer, L. Fernández-Luque, A. Leis, Big Data For Health Through Social Media, Elsevier Inc., 2016. https://doi.org/10.1016/B978-0-12-809269-9.00005-0.

[122] N. Limsopatham, N. Collier, Towards the semantic interpretation of personal health messages from social media, in: UCUI 2015 - Proc. ACM 1st Int. Work. Underst. City with Urban Informatics, Co-Located with CIKM 2015, 2015: pp. 27–30. https://doi.org/10.1145/2811271.2811275.

[123] B.J. Erickson, P. Korfiatis, Z. Akkus, T.L. Kline, Machine learning for medical imaging, Radiographics. 37 (2017) 505–515. https://doi.org/10.1148/rg.2017160130.

[124] S. Ben-david, S. Shalev-shwartz, D. Clarke, R. Schapire, O. Winther, Machine Learning : Foundations and Algorithms, Davidwind.Dk. (2014) 1–14. http://www.davidwind.dk/wp-content/uploads/2014/07/main.pdf (accessed March 6, 2022).

[125] E.G. Dada, J.S. Bassi, H. Chiroma, S.M. Abdulhamid, A.O. Adetunmbi, O.E. Ajibuwa, Machine learning for email spam filtering: review, approaches and open research problems, Heliyon. 5 (2019) e01802. https://doi.org/10.1016/j.heliyon.2019.e01802.

[126] T. Morioka, T. Iwata, T. Hori, T. Kobayashi, Multiscale recurrent neural network based language model, in: Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, 2015: pp. 2366–2370. https://doi.org/10.21437/interspeech.2015-512.

[127] S.J. Mooney, V. Pejaver, Big Data in Public Health: Terminology, Machine Learning, and Privacy, Annu. Rev. Public Health. 39 (2018) 95–112. https://doi.org/10.1146/annurev-publhealth-040617-014208.

[128] X. Dai, M. Bikdash, B. Meyer, From social media to public health surveillance: Word embedding based clustering method for twitter classification, Conf. Proc. - IEEE SOUTHEASTCON. (2017) 1–7. https://doi.org/10.1109/SECON.2017.7925400.

[129] F. Godin, B. Vandersmissen, W. De Neve, R. Van De Walle, Multimedia Lab @ ACL W-NUT

NER Shared Task: Named Entity Recognition for Twitter Microposts using Distributed Word Representations, 2015.

[130] L.P. Kaelbling, M.L. Littman, A.W. Moore, Reinforcement learning: A survey, J. Artif. Intell. Res. 4 (1996) 237–285. https://doi.org/10.1613/jair.301.

[131] K. Kalaiselvi, D. Karthika, Identifying Diseases and Diagnosis Using Machine Learning, in: Springer, Cham, 2020: pp. 391–415. https://doi.org/10.1007/978-3-030-40850-3_16.

[132] Project Hanover, Microsoft Res. (2016). https://www.microsoft.com/en-us/research/project/project-hanover/ (accessed April 6, 2022).

[133] S. Pandya, A. Thakur, S. Saxena, N. Jassal, C. Patel, K. Modi, P. Shah, R. Joshi, S. Gonge, K. Kadam, P. Kadam, A study of the recent trends of immunology: Key challenges, domains, applications, datasets, and future directions, Sensors. 21 (2021) 7786. https://doi.org/10.3390/s21237786.

[134] S. Turgeon, M.J. Lanovaz, Tutorial: Applying Machine Learning in Behavioral Research, Perspect. Behav. Sci. 43 (2020) 697–723. https://doi.org/10.1007/s40614-020-00270-y.

[135] M. Thomas, 15 Machine Learning in Healthcare Examples to Know, Built In. (2020). https://builtin.com/artificial-intelligence/machine-learning-healthcare (accessed April 4, 2022).

[136] M. Field, N. Hardcastle, M. Jameson, N. Aherne, L. Holloway, Machine learning applications in radiation oncology, Phys. Imaging Radiat. Oncol. 19 (2021) 13–24. https://doi.org/10.1016/j.phro.2021.05.007.

[137] Q. Shao, Y. Xu, H. Wu, Spatial Prediction of COVID-19 in China Based on Machine Learning Algorithms and Geographically Weighted Regression, Comput. Math. Methods Med. 2021 (2021). https://doi.org/10.1155/2021/7196492.

[138] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022. https://doi.org/10.1016/b978-0-12-411519-4.00006-9.

[139] D. Zeng, Z. Cao, D.B. Neill, Artificial intelligence–enabled public health surveillance—from local detection to global epidemic monitoring and control, in: Artif. Intell. Med., Elsevier, 2021: pp. 437–453. https://doi.org/10.1016/b978-0-12-821259-2.00022-3.

[140] N.L. Cobb, N.A. Sathe, K.I. Duan, K.P. Seitz, M.R. Thau, C.C. Sung, E.D. Morrell, C. Mikacenic, H.N. Kim, W.C. Liles, A.M. Luks, J. Town, S. Pipavath, M.M. Wurfel, C.L. Hough, T.E. West, P.K. Bhatraju, Comparison of clinical features and outcomes in critically Ill patients hospitalized with COVID-19 versus influenza, Ann. Am. Thorac. Soc. 18 (2021) 632–640. https://doi.org/10.1513/AnnalsATS.202007-805OC.

[141] F. Jaotombo, V. Pauly, P. Auquier, V. Orleans, M. Boucekine, G. Fond, B. Ghattas, L. Boyer, Machine-learning prediction of unplanned 30-day rehospitalization using the French hospital

medico-administrative database, Medicine (Baltimore). 99 (2020) e22361. https://doi.org/10.1097/MD.0000000000022361.

[142] C.C. John, V. Ponnusamy, S. Krishnan Chandrasekaran, N. Ra, A Survey on Mathematical, Machine Learning and Deep Learning Models for COVID-19 Transmission and Diagnosis, IEEE Rev. Biomed. Eng. 15 (2022) 325–340. https://doi.org/10.1109/RBME.2021.3069213.

[143] S. Hong, Y.S. Liu, B. Cao, J. Cao, M. Ai, J. Chen, A. Greenshaw, L. Kuang, Identification of suicidality in adolescent major depressive disorder patients using sMRI: A machine learning approach., J. Affect. Disord. 280 (2021) 72–76. https://doi.org/10.1016/j.jad.2020.10.077.

[144] M. Yang, M. Kiang, W. Shang, Filtering big data from social media - Building an early warning system for adverse drug reactions, J. Biomed. Inform. 54 (2015) 230–240. https://doi.org/10.1016/j.jbi.2015.01.011.

[145] K.W. Johnson, J. Torres Soto, B.S. Glicksberg, K. Shameer, R. Miotto, M. Ali, E. Ashley, J.T. Dudley, Artificial Intelligence in Cardiology, J. Am. Coll. Cardiol. 71 (2018) 2668–2679. https://doi.org/10.1016/j.jacc.2018.03.521.

[146] A. Rushdy, M. O'Mahony, PHLS overview of communicable diseases 1997: results of a priority setting exercise., Commun. Dis. Rep. CDR Suppl. 8 (1998).

[147] R.S. Hopkins, Design and operation of state and local infectious disease surveillance systems, J. Public Heal. Manag. Pract. 11 (2005) 184–190. https://doi.org/10.1097/00124784-200505000-00002.

[148] World Health Organization Emerging and other Communicable Diseases, Surveillance and Control, 1996. http://www.who.int/emc (accessed March 24, 2022).

[149] J.A. Doherty, Establishing priorities for national communicable disease surveillance, Can. J. Infect. Dis. 11 (2000) 21–24. https://doi.org/10.1155/2000/134624.

[150] Hannah Ritchie, Max Roser, Causes of Death - Our World in Data, Our World Data. (2019). https://ourworldindata.org/causes-of-death (accessed March 24, 2022).

[151] Coronavirus now the leading cause of death globally, (n.d.). https://www.thenews.com.pk/latest/676707-coronavirus-now-the-leading-cause-of-death-globally (accessed February 16, 2022).

[152] Just How Do Deaths Due to COVID-19 Stack Up? | Think Global Health, (n.d.). https://www.thinkglobalhealth.org/article/just-how-do-deaths-due-covid-19-stack (accessed March 24, 2022).

[153] Pandey Kiran, COVID-19 now 2nd biggest cause of death in India, Down To Earth. (2021). https://www.downtoearth.org.in/news/health/covid-19-now-2nd-biggest-cause-of-death-in-india-76752 (accessed March 24, 2022).

[154] K. Lee, A. Agrawal, A. Choudhary, Real-Time disease surveillance using twitter data:Demonstration on flu and cancer, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2013: pp. 1474–1477. https://doi.org/10.1145/2487575.2487709.

[155] I.C.-H. Fung, Z.T.H. Tse, K.-W. Fu, The use of social media in public health surveillance, 6 (2015) 10–13. https://doi.org/10.5365/wpsar.2015.6.1.019.

[156] E. Boonchieng, K. Duangchaemkarn, Digital disease detection: Application of machine learning in community health informatics, 2016 13th Int. Jt. Conf. Comput. Sci. Softw. Eng. JCSSE 2016. (2016) 1–5. https://doi.org/10.1109/JCSSE.2016.7748841.

[157] M. Bates, Tracking Disease: Digital Epidemiology Offers New Promise in Predicting Outbreaks, IEEE Pulse. 8 (2017) 18–22. https://doi.org/10.1109/MPUL.2016.2627238.

[158] S. Saini, S. Kohli, Machine Learning Techniques for Effective Text Analysis of Social Network E-health Data, (2016) 3783–3788.

[159] V.K. Jain, S. Kumar, Effective surveillance and predictive mapping of mosquito-borne diseases using social media, J. Comput. Sci. 25 (2018) 406–415. https://doi.org/10.1016/j.jocs.2017.07.003.

[160] V.K. Jain, S. Kumar, An Effective Approach to Track Levels of Influenza-A (H1N1) Pandemic in India Using Twitter, in: Procedia Comput. Sci., Elsevier Masson SAS, 2015: pp. 801–807. https://doi.org/10.1016/j.procs.2015.10.120.

[161] W. Zhang, S. Ram, M. Burkart, Y. Pengetnze, Extracting Signals from Social Media for Chronic Disease Surveillance, (2016) 79–83. https://doi.org/10.1145/2896338.2897728.

[162] S. Wakamiya, Y. Kawai, E. Aramaki, After the boom no one tweets: Microblog-based influenza detection incorporating indirect information, in: ACM Int. Conf. Proceeding Ser., 2016: pp. 17–25. https://doi.org/10.1145/3007818.3007822.

[163] X. Ji, S.A. Chun, J. Geller, Monitoring public health concerns using twitter sentiment classifications, Proc. - 2013 IEEE Int. Conf. Healthc. Informatics, ICHI 2013. (2013) 335–344. https://doi.org/10.1109/ICHI.2013.47.

[164] C. Allen, M.H. Tsou, A. Aslam, A. Nagel, J.M. Gawron, Applying GIS and machine learning methods to twitter data for multiscale surveillance of influenza, PLoS One. 11 (2016) 1–10. https://doi.org/10.1371/journal.pone.0157734.

[165] K.M. Han J, Data Mining : Concepts and Techniques. San Francisco: Morgan Kaufmann Publisher, 2006.

[166] K. Koutroumbas, N. Kalouptsidis, Nearest neighbor pattern classification neural networks, (2002) 2911–2915. https://doi.org/10.1109/icnn.1994.374694.

[167] K. Jiang, R. Gupta, M. Gupta, R.A. Calix, G.R. Bernard, Identifying Personal Health Experience Tweets with Deep Neural Networks* HHS Public Access, Conf Proc IEEE Eng Med Biol Soc.

2017 (2017) 1174–1177. https://doi.org/10.1109/EMBC.2017.8037039.

[168] C.Y.J. Peng, K.L. Lee, G.M. Ingersoll, An introduction to logistic regression analysis and reporting, J. Educ. Res. 96 (2002) 3–14. https://doi.org/10.1080/00220670209598786.

[169] I. Korkontzelos, D. Piliouras, A.W. Dowsey, S. Ananiadou, Boosting drug named entity recognition using an aggregate classifier, Artif. Intell. Med. 65 (2015) 145–153. https://doi.org/10.1016/j.artmed.2015.05.007.

[170] L. Zhao, J. Chen, F. Chen, W. Wang, C.T. Lu, N. Ramakrishnan, SimNest: Social media nested epidemic simulation via online semi-supervised deep learning, Proc. - IEEE Int. Conf. Data Mining, ICDM. 2016-Janua (2016) 639–648. https://doi.org/10.1109/ICDM.2015.39.

[171] S. Rasoul Safavian, D. Landgrebe, A Survey of Decision Tree Classifier Methodology, (2017).

[172] S. Chae, S. Kwon, D. Lee, Predicting infectious disease using deep learning and big data, Int. J. Environ. Res. Public Health. 15 (2018). https://doi.org/10.3390/ijerph15081596.

[173] M.A. Carlos, M. Nogueira, R.J. Machado, Analysis of dengue outbreaks using big data analytics and social networks, 2017 4th Int. Conf. Syst. Informatics, ICSAI 2017. 2018-Janua (2018) 1592–1597. https://doi.org/10.1109/ICSAI.2017.8248538.

[174] P. Kostkova, S. Garbin, J. Moser, W. Pan, Integration and visualization public health dashboard, in: Proc. 23rd Int. Conf. World Wide Web - WWW '14 Companion, 2014: pp. 657–662. http://dl.acm.org/citation.cfm?id=2567948.2579276&coll=DL&dl=GUIDE&CFID=574099975&CFTOKEN=12201186%0Ahttp://dl.acm.org/citation.cfm?doid=2567948.2579276.

[175] S. Tuarob, C.S. Tucker, M. Salathe, N. Ram, Discovering health-related knowledge in social media using ensembles of heterogeneous features, in: Int. Conf. Inf. Knowl. Manag. Proc., 2013: pp. 1685–1690. https://doi.org/10.1145/2505515.2505629.

[176] H. Abbasimehr, R. Paki, Prediction of COVID-19 confirmed cases combining deep learning methods and Bayesian optimization, Chaos, Solitons and Fractals. 142 (2021). https://doi.org/10.1016/j.chaos.2020.110511.

[177] L. Luo, Y. Wang, H. Liu, COVID-19 personal health mention detection from tweets using dual convolutional neural network, Expert Syst. Appl. 200 (2022). https://doi.org/10.1016/j.eswa.2022.117139.

[178] M.M. Tadesse, H. Lin, B. Xu, L. Yang, Detection of suicide ideation in social media forums using deep learning, Algorithms. 13 (2020) 7. https://doi.org/10.3390/a13010007.

[179] K. Zeberga, M. Attique, B. Shah, F. Ali, Y.Z. Jembre, T.S. Chung, A Novel Text Mining Approach for Mental Health Prediction Using Bi-LSTM and BERT Model, Comput. Intell. Neurosci. 2022 (2022). https://doi.org/10.1155/2022/7893775.

[180] J.A. Benitez-Andrades, J.M. Alija-Perez, I. Garcia-Rodriguez, C. Benavides, H. Alaiz-Moreton,

R.P. Vargas, M.T. Garcia-Ordas, BERT model-based approach for detecting categories of tweets in the field of eating disorders (ED), in: Proc. - IEEE Symp. Comput. Med. Syst., 2021: pp. 586–590. https://doi.org/10.1109/CBMS52027.2021.00105.

[181] M. Kamal, S.U. Rehman khan, S. Hussain, A. Nasir, K. Aslam, S. Tariq, M.F. Ullah, Predicting mental illness using social media posts and comments, Int. J. Adv. Comput. Sci. Appl. 11 (2020) 607–613. https://doi.org/10.14569/IJACSA.2020.0111271.

[182] P. Kumar, P. Samanta, S. Dutta, M. Chatterjee, D. Sarkar, Feature Based Depression Detection from Twitter Data Using Machine Learning Techniques, J. Sci. Res. 66 (2022) 220–228. https://doi.org/10.37398/jsr.2022.660229.

[183] J. Kim, J. Lee, E. Park, J. Han, A deep learning model for detecting mental illness from user content on social media, Sci. Rep. 10 (2020) 1–6. https://doi.org/10.1038/s41598-020-68764-y.

[184] S. Hochreiter, J Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780. https://ieeexplore.ieee.org/abstract/document/6795963/ (accessed September 19, 2021).

[185] E.K. Priya Sri, K.S. Savita, M. Zaffar, Depression Detection in Tweets from Urban Cities of Malaysia using Deep Learning, in: Int. Conf. Res. Innov. Inf. Syst. ICRIIS, 2021. https://doi.org/10.1109/ICRIIS53035.2021.9617079.

[186] R. Biddle, A. Joshi, S. Liu, C. Paris, G. Xu, Leveraging Sentiment Distributions to Distinguish Figurative from Literal Health Reports on Twitter, in: Web Conf. 2020 - Proc. World Wide Web Conf. WWW 2020, Association for Computing Machinery, Inc, 2020: pp. 1217–1227. https://doi.org/10.1145/3366423.3380198.

[187] J. Du, Y. Zhang, J. Luo, Y. Jia, Q. Wei, C. Tao, H. Xu, Extracting psychiatric stressors for suicide from social media using deep learning, BMC Med. Inform. Decis. Mak. 18 (2018). https://doi.org/10.1186/s12911-018-0632-8.

[188] R. Sawhney, R.R. Shah, V. Bhatia, C.T. Lin, S. Aggarwal, M. Prasad, Exploring the Impact of Evolutionary Computing based Feature Selection in Suicidal Ideation Detection, in: IEEE Int. Conf. Fuzzy Syst., Institute of Electrical and Electronics Engineers Inc., 2019. https://doi.org/10.1109/FUZZ-IEEE.2019.8858989.

[189] S. Wang, J. Du, L. Tang, C. Tao, Understanding Public Perceptions of Measles from Twitter Using Multi-Task Convolutional Neural Networks, in: Stud. Health Technol. Inform., Stud Health Technol Inform, 2022. https://doi.org/10.3233/shti220149.

[190] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Adv. Neural Inf. Process. Syst., 2017: pp. 5999–6009. https://www.autoriteitnvs.nl/onderwerpen/stralingsincidenten/inhoud/incidenten-met-a--en-of-b-objecten (accessed July 21, 2022).

[191] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., Association for Computational Linguistics (ACL), 2019: pp. 4171–4186. https://doi.org/10.48550/arxiv.1810.04805.

[192] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, Arxiv.Org. (2019). https://arxiv.org/abs/1907.11692 (accessed July 20, 2022).

[193] Huggingface.co, GitHub - huggingface/transformers: ? Transformers: State-of-the-art Machine Learning for Pytorch, TensorFlow, and JAX., (2021). https://github.com/huggingface/transformers (accessed July 20, 2022).

[194] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, (2019). http://arxiv.org/abs/1910.01108 (accessed July 20, 2022).

[195] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, XLNet: Generalized autoregressive pretraining for language understanding, in: Adv. Neural Inf. Process. Syst., Neural information processing systems foundation, 2019. https://doi.org/10.48550/arxiv.1906.08237.

[196] A. Murarka, I.B.M. Raleigh, Classification of mental illnesses on social media using RoBERTa, in: Proc. Ofthe 12th Int. Work. Heal. Text Min. Inf. Anal., 2021: pp. 59–68. https://aclanthology.org/2021.louhi-1.7/ (accessed July 20, 2022).

[197] X. Wang, S. Chen, T. Li, W. Li, Y. Zhou, J. Zheng, Q. Chen, J. Yan, B. Tang, Depression risk prediction for chinese microblogs via deep-learning methods: Content analysis, JMIR Med. Informatics. 8 (2020) e17958. https://doi.org/10.2196/17958.

[198] W. 0 Kermack, A.G. Mckendrick, A contribution to the mathematical theory of epidemics, Proc. R. Soc. London. Ser. A, Contain. Pap. a Math. Phys. Character. 115 (1927) 700–721. https://doi.org/10.1098/rspa.1927.0118.

[199] G. Chowell, A. Tariq, J.M. Hyman, A novel sub-epidemic modeling framework for short-term forecasting epidemic waves, BMC Med. 17 (2019) 1–18. https://doi.org/10.1186/s12916-019-1406-6.

[200] F.J. Richards, A flexible growth function for empirical use, J. Exp. Bot. 10 (1959) 290–301. https://doi.org/10.1093/jxb/10.2.290.

[201] G. Chowell, Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts, Infect. Dis. Model. 2 (2017) 379–398. https://doi.org/10.1016/j.idm.2017.08.001.

[202] S. Saha, G.P. Samanta, J.J. Nieto, Epidemic model of COVID-19 outbreak by inducing behavioural response in population, Nonlinear Dyn. 102 (2020) 455–487.

https://doi.org/10.1007/s11071-020-05896-w.

[203] B. Ambikapathy, K. Krishnamurthy, Mathematical modelling to assess the impact of lockdown on COVID-19 transmission in India: Model development and validation, JMIR Public Heal. Surveill. 6 (2020) e19368. https://doi.org/10.2196/19368.

[204] P. Shao, Impact of city and residential unit lockdowns on prevention and control of COVID-19, MedRxiv. (2020). https://doi.org/10.1101/2020.03.13.20035253.

[205] K. Akinci, J. Fdez, E. Peña-Tapia, O. Witkowski, Dynamic versus Continuous Interventions: Optimizing Lockdown Policies for COVID-19, MedRxiv. (2021) 2021.03.10.21253324. https://doi.org/10.1101/2021.03.10.21253324.

[206] X. Li, X. Wu, Long short-term memory based convolutional recurrent neural networks for large vocabulary speech recognition, in: Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, 2015: pp. 3219–3223. https://doi.org/10.21437/interspeech.2015-648.

[207] Z. Du, X. Lin, Air Quality Prediction Based on Neural Network Model of Long Short-term Memory, in: IOP Conf. Ser. Earth Environ. Sci., IOP Publishing, 2020: p. 012013. https://doi.org/10.1088/1755-1315/508/1/012013.

[208] A. Gupta, R. Katarya, PAN-LDA: A latent Dirichlet allocation based novel feature extraction model for COVID-19 data using machine learning, Comput. Biol. Med. 138 (2021) 104920. https://doi.org/10.1016/j.compbiomed.2021.104920.

[209] N. Zheng, S. Du, J. Wang, H. Zhang, W. Cui, Z. Kang, T. Yang, B. Lou, Y. Chi, H. Long, M. Ma, Q. Yuan, S. Zhang, D. Zhang, F. Ye, J. Xin, Predicting COVID-19 in China Using Hybrid AI Model, IEEE Trans. Cybern. 50 (2020) 2891–2904. https://doi.org/10.1109/TCYB.2020.2990162.

[210] N. Imai, A. Cori, I. Dorigatti, M. Baguelin, C.A. Donnelly, S. Riley, N.M. Ferguson, Transmissibility of 2019-nCoV, World Heal. Organ. (2019) 2–6.

[211] Z. Liu, P. Magal, O. Seydi, G. Webb, Predicting the cumulative number of cases for the COVID-19 epidemic in China from early data, Math. Biosci. Eng. 17 (2020) 3040–3051. https://doi.org/10.3934/MBE.2020172.

[212] A. Gupta, R. Katarya, Social media based surveillance systems for healthcare using machine learning: A systematic review, J. Biomed. Inform. 108 (2020) 103500–103500. https://doi.org/10.1016/j.jbi.2020.103500.

[213] World Health Organization, India: WHO Coronavirus Disease (COVID-19) Dashboard With Vaccination Data, World Heal. Organ. (2021) 1–5. https://covid19.who.int/region/searo/country/in (accessed September 21, 2021).

[214] Government of India, COVID-19 in India, Vaccination, Dashboard , Corona Virus Tracker | mygov.in, Gov. India. (2021). https://www.mygov.in/covid-19 (accessed September 21, 2021).

[215] M. Souris, J.P. Gonzalez, COVID-19: Spatial analysis of hospital casefatality rate in France, PLoS One. 15 (2020) e0243606. https://doi.org/10.1371/journal.pone.0243606.

[216] G. Zhang, X. Liu, Prediction and control of COVID-19 spreading based on a hybrid intelligent model, PLoS One. 16 (2021) e0246360. https://doi.org/10.1371/journal.pone.0246360.

[217] L. Ang, H.W. Lee, A. Kim, M.S. Lee, Herbal medicine for the management of COVID-19 during the medical observation period: a review of guidelines, Integr. Med. Res. 9 (2020) 100465. https://doi.org/10.1016/j.imr.2020.100465.

[218] J.A. Backer, D. Klinkenberg, J. Wallinga, Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20 28 January 2020, Eurosurveillance. 25 (2020) 20–28. https://doi.org/10.2807/1560-7917.ES.2020.25.5.2000062.

[219] H. Meisheri, K. Ranjan, L. Dey, Sentiment Extraction from Consumer-Generated Noisy Short Texts, in: IEEE Int. Conf. Data Min. Work. ICDMW, 2017: pp. 399–406. https://doi.org/10.1109/ICDMW.2017.58.

[220] D.P. Kingma, J.L. Ba, Adam: A method for stochastic optimization, in: 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., International Conference on Learning Representations, ICLR, 2015.

[221] S. Du, J. Wang, H. Zhang, W. Cui, Z. Kang, T. Yang, B. Lou, Y. Chi, H. Long, M. Ma, Q. Yuan, S. Zhang, D. Zhang, J. Xin, N. Zheng, Predicting COVID-19 Using Hybrid AI Model, SSRN Electron. J. (2020). https://doi.org/10.2139/ssrn.3555202.

[222] S. Metelmann, K. Pattni, L. Brierley, L. Cavalerie, C. Caminade, M.S.C. Blagrove, J. Turner, K.J. Sharkey, M. Baylis, Impact of climatic, demographic and disease control factors on the transmission dynamics of COVID-19 in large cities worldwide, One Heal. 12 (2021). https://doi.org/10.1016/j.onehlt.2021.100221.

[223] J. Duhon, N. Bragazzi, J.D. Kong, The impact of non-pharmaceutical interventions, demographic, social, and climatic factors on the initial growth rate of COVID-19: A cross-country study, Sci. Total Environ. 760 (2021) 144325. https://doi.org/10.1016/j.scitotenv.2020.144325.

[224] D.R. Petretto, R. Pili, Ageing and COVID-19: What Is the Role for Elderly People?, Geriatrics. 5 (2020) 25. https://doi.org/10.3390/geriatrics5020025.

[225] G. Lippi, B.M. Henry, C. Mattiuzzi, C. Bovo, The death rate for covid-19 is positively associated with gross domestic products, Acta Biomed. 91 (2020) 224–225. https://doi.org/10.23750/abm.v91i2.9514.

[226] C. Copat, A. Cristaldi, M. Fiore, A. Grasso, P. Zuccarello, S.S. Signorelli, G.O. Conti, M. Ferrante, The role of air pollution (PM and NO2) in COVID-19 spread and lethality: A systematic review, Environ. Res. 191 (2020) 110129. https://doi.org/10.1016/j.envres.2020.110129.

[227] R. Tosepu, J. Gunawan, D.S. Effendy, L.O.A.I. Ahmad, H. Lestari, H. Bahar, P. Asfian, Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia, Sci. Total Environ. 725 (2020) 138436. https://doi.org/10.1016/j.scitotenv.2020.138436.

[228] A. Bhadra, A. Mukherjee, K. Sarkar, Impact of population density on Covid-19 infected and mortality rate in India, Model. Earth Syst. Environ. 7 (2021) 623–629. https://doi.org/10.1007/s40808-020-00984-7.

[229] N. Imai, K.A.M. Gaythorpe, S. Abbott, S. Bhatia, S. van Elsland, K. Prem, Y. Liu, N.M. Ferguson, Adoption and impact of non-pharmaceutical interventions for COVID-19, Wellcome Open Res. 5 (2020). https://doi.org/10.12688/wellcomeopenres.15808.1.

[230] PRS, PRS India, COVID-19 Cases, (n.d.). https://prsindia.org/covid-19/cases.

[231] TuTiempo.net, World Weather - Local Weather Forecast, (n.d.). https://en.tutiempo.net/ (accessed March 9, 2021).

[232] Census 2011, Density of India State of Population Census 2011, (n.d.). https://www.census2011.co.in/density.php (accessed March 9, 2021).

[233] Ministry of Statistics and Program Implementation, Research Studies Comparative | Ministry of Statistics and Program Implementation | Government Of India, (n.d.). http://mospi.nic.in/data (accessed March 9, 2021).

[234] RBI, Handbook of Statistics on Indian States Reserve Bank of India 2019-20, 2020. https://m.rbi.org.in/Scripts/AnnualPublications.aspx?head=Handbook+of+Statistics+on+Indian+States (accessed March 9, 2021).

[235] COVID-19 - Center for Disease Dynamics, Economics & Policy (CDDEP), (n.d.). https://cddep.org/research-area/covid-19/ (accessed September 22, 2021).

[236] R. Haberman, Mathematical Models: Mechanical Vibrations, Population Dynamics, and Traffic Flow, Society for Industrial and Applied Mathematics, 1998. https://doi.org/10.1137/1.9781611971156.

[237] X.S. Wang, J. Wu, Y. Yang, Richards model revisited: Validation by and application to infection dynamics, J. Theor. Biol. 313 (2012) 12–19. https://doi.org/10.1016/j.jtbi.2012.07.024.

[238] H. Zhang, L. Yang, L. Li, G. Xu, X. Zhang, The epidemic characteristics and spatial autocorrelation analysis of hand, foot and mouth disease from 2010 to 2015 in Shantou, Guangdong, China, BMC Public Health. 19 (2019) 998. https://doi.org/10.1186/s12889-019-7329-5.

[239] R.M.F. de Arruda, D.T. Cardoso, R.G. Teixeira-Neto, D.S. Barbosa, R.K. Ferraz, M.H.F. Morais, V.S. Belo, E.S. da Silva, Space-time analysis of the incidence of human visceral leishmaniasis (VL) and prevalence of canine VL in a municipality of southeastern Brazil: Identification of

priority areas for surveillance and control, Acta Trop. 197 (2019) 105052. https://doi.org/10.1016/j.actatropica.2019.105052.

[240] L. Anselin, Local Indicators of Spatial Association—LISA, Geogr. Anal. 27 (1995) 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

[241] T. Monzur, Local Spatial Autocorrelation - 1, (2005). https://doi.org/10.13140/RG.2.1.4708.0404.

[242] Q. Xu, Y. Dong, Y. Wang, R. Yang, C. Xu, Determinants and identification of the northern boundary of China's tropical zone, J. Geogr. Sci. 28 (2018) 31–45. https://doi.org/10.1007/s11442-018-1457-1.

[243] J. Wang, C. Xu, Geodetector: Principle and prospective, Dili Xuebao/Acta Geogr. Sin. 72 (2017) 116–134. https://doi.org/10.11821/dlxb201701010.

[244] Y. Han, W. Zhao, P. Pereira, Global COVID-19 pandemic trends and their relationship with meteorological variables, air pollutants and socioeconomic aspects, Environ. Res. 204 (2022) 112249. https://doi.org/10.1016/j.envres.2021.112249.

[245] CDC, Science Brief: SARS-CoV-2 and Potential Airborne Transmission. Centers for Disease Control and Prevention, (2020). https://www.cdc.gov/coronavirus/2019-ncov/more/scientific-brief-sars-cov-2.html (accessed March 11, 2021).

[246] J.F.W. Chan, S. Yuan, K.H. Kok, K.K.W. To, H. Chu, J. Yang, F. Xing, J. Liu, C.C.Y. Yip, R.W.S. Poon, H.W. Tsoi, S.K.F. Lo, K.H. Chan, V.K.M. Poon, W.M. Chan, J.D. Ip, J.P. Cai, V.C.C. Cheng, H. Chen, C.K.M. Hui, K.Y. Yuen, A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster, Lancet. 395 (2020) 514–523. https://doi.org/10.1016/S0140-6736(20)30154-9.

[247] N. Kadi, M. Khelfaoui, Population density, a factor in the spread of COVID-19 in Algeria: statistic study, Bull. Natl. Res. Cent. 44 (2020) 1–7. https://doi.org/10.1186/s42269-020-00393-x.

[248] S. Hamidi, S. Sabouri, R. Ewing, Does Density Aggravate the COVID-19 Pandemic?: Early Findings and Lessons for Planners, J. Am. Plan. Assoc. 86 (2020) 495–509. https://doi.org/10.1080/01944363.2020.1777891.

[249] Z. Sun, H. Zhang, Y. Yang, H. Wan, Y. Wang, Impacts of geographic factors and population density on the COVID-19 spreading under the lockdown policies of China, Sci. Total Environ. 746 (2020) 141347. https://doi.org/10.1016/j.scitotenv.2020.141347.

[250] N.D. Yanez, N.S. Weiss, J.A. Romand, M.M. Treggiari, COVID-19 mortality risk for older men and women, BMC Public Health. 20 (2020) 1742. https://doi.org/10.1186/s12889-020-09826-8.

[251] G. Marois, R. Muttarak, S. Scherbov, Assessing the potential impact of COVID-19 on life expectancy, PLoS One. 15 (2020) e0238678. https://doi.org/10.1371/journal.pone.0238678.

[252] R. Paul, A.A. Arif, O. Adeyemi, S. Ghosh, D. Han, Progression of COVID-19 From Urban to

Rural Areas in the United States: A Spatiotemporal Analysis of Prevalence Rates, J. Rural Heal. 36 (2020) 591–601. https://doi.org/10.1111/jrh.12486.

[253] C. El Bcheraoui, A.H. Mokdad, L. Dwyer-Lindgren, A. Bertozzi-Villa, R. WStubbs, C. Morozoff, S. Shirude, M. Naghavi, C.J.L. Murray, Trends and patterns of differences in infectious disease mortality among US Counties, 1980-2014, JAMA - J. Am. Med. Assoc. 319 (2018) 1248–1260. https://doi.org/10.1001/jama.2018.2089.

[254] M.M. Menebo, Temperature and precipitation associate with Covid-19 new daily cases: A correlation study between weather and Covid-19 pandemic in Oslo, Norway, Sci. Total Environ. 737 (2020) 139659. https://doi.org/10.1016/j.scitotenv.2020.139659.

[255] P. Mecenas, R.T. da R.M. Bastos, A.C.R. Vallinoto, D. Normando, Effects of temperature and humidity on the spread of COVID-19: A systematic review, PLoS One. 15 (2020) e0238339. https://doi.org/10.1371/journal.pone.0238339.

[256] R. Li, C. Rivers, Q. Tan, M.B. Murray, E. Toner, M. Lipsitch, Estimated Demand for US Hospital Inpatient and Intensive Care Unit Beds for Patients With COVID-19 Based on Comparisons With Wuhan and Guangzhou, China, JAMA Netw. Open. 3 (2020) e208297. https://doi.org/10.1001/jamanetworkopen.2020.8297.

[257] P. Karaca-Mandic, S. Sen, A. Georgiou, Y. Zhu, A. Basu, Association of COVID-19-Related Hospital Use and Overall COVID-19 Mortality in the USA, J. Gen. Intern. Med. (2020) 1–3. https://doi.org/10.1007/s11606-020-06084-7.

[258] WHO, India: WHO Coronavirus Disease (COVID-19) Dashboard With Vaccination Data | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data, World Heal. Organ. (2021) 1–5. https://covid19.who.int/region/searo/country/in (accessed October 21, 2021).

[259] COVID19-India API | api, COVID19-India API | api, (2020). https://api.covid19india.org/ (accessed May 31, 2021).

[260] Y. Pan, A. Darzi, A. Kabiri, G. Zhao, W. Luo, C. Xiong, L. Zhang, Quantifying human mobility behaviour changes during the COVID-19 outbreak in the United States, Sci. Rep. 10 (2020) 1–9. https://doi.org/10.1038/s41598-020-77751-2.

[261] Google, Community Mobility Reports, Google. (2021). https://www.google.com/covid19/mobility/ (accessed May 31, 2021).

[262] Weather Underground, Local Weather Forecast, News and Conditions | Weather Underground, (n.d.). https://www.wunderground.com/ (accessed May 31, 2021).

[263] Aqicn.org, Air Pollution in India., (n.d.). https://aqicn.org/city/india (accessed September 1, 2021).

[264] V. Moorthy, A.M.H. Restrepo, M.P. Preziosi, S. Swaminathan, Data sharing for novel coronavirus (COVID-19), Bull. World Health Organ. 98 (2020) 150. https://doi.org/10.2471/BLT.20.251561.

[265] K. Wu, D. Darcet, Q. Wang, D. Sornette, Generalized logistic growth modeling of the COVID-19 outbreak: comparing the dynamics in the 29 provinces in China and in the rest of the world, Nonlinear Dyn. 101 (2020) 1561–1581. https://doi.org/10.1007/s11071-020-05862-6.

[266] A.L. Bertozzi, E. Franco, G. Mohler, M.B. Short, D. Sledge, The challenges of modeling and forecasting the spread of COVID-19, Proc. Natl. Acad. Sci. U. S. A. 117 (2020) 16732–16738. https://doi.org/10.1073/pnas.2006520117.

[267] W. Dhouib, J. Maatoug, I. Ayouni, N. Zammit, R. Ghammem, S. Ben Fredj, H. Ghannem, The incubation period during the pandemic of COVID-19: a systematic review and meta-analysis, Syst. Rev. 10 (2021) 1–14. https://doi.org/10.1186/s13643-021-01648-y.

[268] J.A. Backer, D. Klinkenberg, J. Wallinga, Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20 28 January 2020, Eurosurveillance. 25 (2020) 2000062. https://doi.org/10.2807/1560-7917.ES.2020.25.5.2000062.

[269] L.N. Nottmeyer, F. Sera, Influence of temperature, and of relative and absolute humidity on COVID-19 incidence in England - A multi-city time-series study, Environ. Res. 196 (2021) 110977. https://doi.org/10.1016/j.envres.2021.110977.

[270] H.A. Rothan, S.N. Byrareddy, The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak, J. Autoimmun. 109 (2020) 102433. https://doi.org/10.1016/j.jaut.2020.102433.

[271] M. Day, Covid-19: identifying and isolating asymptomatic people helped eliminate virus in Italian village, BMJ. 368 (2020) m1165. https://doi.org/10.1136/bmj.m1165.

[272] R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, J. R. Stat. Soc. Ser. B. 58 (1996) 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[273] A.E. Hoerl, R.W. Kennard, Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics. 12 (1970) 55–67. https://doi.org/10.1080/00401706.1970.10488634.

[274] W. Zou, The COVID-19 Pandemic Prediction in the US Based on Machine Learning, in: Proc. - 2020 Int. Conf. Public Heal. Data Sci. ICPHDS 2020, Institute of Electrical and Electronics Engineers Inc., 2020: pp. 283–289. https://doi.org/10.1109/ICPHDS51617.2020.00062.

[275] S.A. Curiskis, B. Drake, T.R. Osborn, P.J. Kennedy, An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit, Inf. Process. Manag. 57 (2020) 102034. https://doi.org/10.1016/j.ipm.2019.04.002.

[276] A. Onan, Two-Stage Topic Extraction Model for Bibliometric Data Analysis Based on Word Embeddings and Clustering, IEEE Access. 7 (2019) 145614–145633. https://doi.org/10.1109/ACCESS.2019.2945911.

[277] G. Salton, A. Wong, C.S. Yang, A Vector Space Model for Automatic Indexing, Commun. ACM.

18 (1975) 613–620. https://doi.org/10.1145/361219.361220.

[278] A. Onan, Biomedical Text Categorization Based on Ensemble Pruning and Optimized Topic Modelling, Comput. Math. Methods Med. 2018 (2018). https://doi.org/10.1155/2018/2497471.

[279] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: 1st Int. Conf. Learn. Represent. ICLR 2013 - Work. Track Proc., International Conference on Learning Representations, ICLR, 2013. http://ronan.collobert.com/senna/ (accessed December 21, 2020).

[280] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Adv. Neural Inf. Process. Syst., Neural information processing systems foundation, 2013.

[281] J. Xie, R. Girshick, A. Farhadi, Unsupervised deep embedding for clustering analysis, in: 33rd Int. Conf. Mach. Learn. ICML 2016, International Machine Learning Society (IMLS), 2016: pp. 740–749. https://arxiv.org/abs/1511.06335v2 (accessed July 25, 2021).

[282] B. Schölkopf, A. Smola, K.R. Müller, Nonlinear Component Analysis as a Kernel Eigenvalue Problem, Neural Comput. 10 (1998) 1299–1319. https://doi.org/10.1162/089976698300017467.

[283] B. Schölkopf, A. Smola, K.R. Müller, Kernel principal component analysis, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), Springer, Berlin, Heidelberg, 1997: pp. 583–588. https://doi.org/10.1007/bfb0020217.

[284] A. Onan, M.A. Tocoglu, A Term Weighted Neural Language Model and Stacked Bidirectional LSTM Based Framework for Sarcasm Identification, IEEE Access. 9 (2021) 7701–7722. https://doi.org/10.1109/ACCESS.2021.3049734.

[285] E.H. Huang, R. Socher, C.D. Manning, A.Y. Ng, Improving word representations via global context and multipleword prototypes, in: 50th Annu. Meet. Assoc. Comput. Linguist. ACL 2012 - Proc. Conf., Association for Computational Linguistics, 2012: pp. 873–882. http://ai.stanford.edu/ (accessed December 21, 2020).

[286] R. Socher, J. Bauer, C.D. Manning, A.Y. Ng, Parsing with compositional vector grammars, in: ACL 2013 - 51st Annu. Meet. Assoc. Comput. Linguist. Proc. Conf., 2013: pp. 455–465. https://aclanthology.org/P13-1045 (accessed January 20, 2022).

[287] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, J. Mach. Learn. Res. 12 (2011) 2493–2537.

[288] A. Onan, Topic-Enriched Word Embeddings for Sarcasm Identification, in: Adv. Intell. Syst. Comput., Springer Verlag, 2019: pp. 293–304. https://doi.org/10.1007/978-3-030-19807-7_29.

[289] X. Rong, word2vec Parameter Learning Explained, (2014). http://arxiv.org/abs/1411.2738 (accessed December 25, 2020).

[290] F. Morin, Y. Bengio, Hierarchical probabilistic neural network language model, in: AISTATS 2005 - Proc. 10th Int. Work. Artif. Intell. Stat., 2005: pp. 246–252.

[291] M.A. Toçoğlu, A. Onan, Sentiment Analysis on Students' Evaluation of Higher Educational Institutions, in: Adv. Intell. Syst. Comput., Springer, Cham, 2021: pp. 1693–1700. https://doi.org/10.1007/978-3-030-51156-2_197.

[292] M. Steinbach, L. Ertöz, V. Kumar, The Challenges of Clustering High Dimensional Data, in: New Dir. Stat. Phys., 2004: pp. 273–309. https://doi.org/10.1007/978-3-662-08968-2_16.

[293] M. James and others, Some methods for classification and analysis of multivariate observations, Proc. Fifth Berkeley Symp. Math. Stat. Probab. 1 (1967) 281–297. https://projecteuclid.org/euclid.bsmsp/1200512992 (accessed September 8, 2020).

[294] A.K. Jain, Data Clustering: 50 Years Beyond K-means, in: Mach. Learn. Knowl. Discov. Databases, Springer Berlin Heidelberg, 2008: pp. 3–4. https://doi.org/10.1007/978-3-540-87479-9_3.

[295] A. Onan, Hybrid supervised clustering based ensemble scheme for text classification, Kybernetes. 46 (2017) 330–348. https://doi.org/10.1108/K-10-2016-0300.

[296] A. Huang, Similarity measures for text document clustering, in: New Zeal. Comput. Sci. Res. Student Conf. NZCSRSC 2008 - Proc., 2008: pp. 49–56.

[297] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature. 323 (1986) 533–536. https://doi.org/10.1038/323533a0.

[298] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature. 521 (2015) 436–444. https://doi.org/10.1038/nature14539.

[299] S. Abirami, P. Chitra, Energy-efficient edge based real-time healthcare support system, in: Adv. Comput., Elsevier, 2020: pp. 339–368. https://doi.org/10.1016/bs.adcom.2019.09.007.

[300] D.M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, S.S. Ghosh, Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during COVID-19: Observational study, J. Med. Internet Res. 22 (2020). https://doi.org/10.2196/22635.

[301] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., IEEE Computer Society, 2015: pp. 1–9. https://doi.org/10.1109/CVPR.2015.7298594.

[302] L. Van Der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2625.

[303] E. Amigó, J. Gonzalo, F. Verdejo, D. Spina, A comparison of filtering evaluation metrics based on formal constraints, Inf. Retr. J. 22 (2019) 581–619. https://doi.org/10.1007/s10791-019-09355-y.

[304] M.P. Naik, H.B. Prajapati, V.K. Dabhi, A survey on semantic document clustering, in: Proc. 2015 IEEE Int. Conf. Electr. Comput. Commun. Technol. ICECCT 2015, Institute of Electrical and Electronics Engineers Inc., 2015. https://doi.org/10.1109/ICECCT.2015.7226036.

[305] L. The Vinh, S. Lee, Y.-T. Park, A novel feature selection method based on normalized mutual information, Appl Intell. 37 (2012) 100–120. https://doi.org/10.1007/s10489-011-0315-y.

[306] J.M. Santos, M. Embrechts, On the use of the adjusted rand index as a metric for evaluating supervised classification, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2009: pp. 175–184. https://doi.org/10.1007/978-3-642-04277-5_18.

[307] A. Amelio, C. Pizzuti, Correction for Closeness: Adjusting Normalized Mutual Information Measure for Clustering Comparison, Comput. Intell. 33 (2017) 579–601. https://doi.org/10.1111/coin.12100.

[308] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1985) 193–218. https://doi.org/10.1007/BF01908075.

[309] A. Tissaoui, S. Sassi, R. Chbeir, Probabilistic Topic Models for Enriching Ontology from Texts, SN Comput. Sci. 1 (2020). https://doi.org/10.1007/s42979-020-00349-y.

[310] X. Li, L. Lei, A bibliometric analysis of topic modelling studies (2000–2017), J. Inf. Sci. 47 (2021) 161–175. https://doi.org/10.1177/0165551519877049.

[311] B. Zhu, X. Zheng, H. Liu, J. Li, P. Wang, Analysis of spatiotemporal characteristics of big data on social media sentiment with COVID-19 epidemic topics, Chaos, Solitons and Fractals. 140 (2020) 110123. https://doi.org/10.1016/j.chaos.2020.110123.

[312] C. Ordun, S. Purushotham, E. Raff, Exploratory Analysis of Covid-19 Tweets using Topic Modeling, UMAP, and DiGraphs, ArXiv. (2020). https://radimrehurek.com/gensim/models/ldamulticore.html (accessed November 27, 2020).

[313] A. Rortais, F. Barrucci, V. Ercolano, J. Linge, A. Christodoulidou, J.P. Cravedi, R. Garcia-Matas, C. Saegerman, L. Svečnjak, A topic model approach to identify and track emerging risks from beeswax adulteration in the media, Food Control. 119 (2021) 107435. https://doi.org/10.1016/j.foodcont.2020.107435.

[314] T. Chuluunsaikhan, G.A. Ryu, K.H. Yoo, H. Rah, A. Nasridinov, Incorporating deep learning and news topic modeling for forecasting pork prices: The case of South Korea, Agric. 10 (2020) 1–22. https://doi.org/10.3390/agriculture10110513.

[315] X. Li, W. Shang, S. Wang, Text-based crude oil price forecasting: A deep learning approach, Int. J. Forecast. 35 (2019) 1548–1560. https://doi.org/10.1016/j.ijforecast.2018.07.006.

[316] A. Mahadevan, M. Arock, Integrated topic modeling and sentiment analysis: A review rating prediction approach for recommender systems, Turkish J. Electr. Eng. Comput. Sci. 28 (2020)

107–123. https://doi.org/10.3906/elk-1905-114.

[317] A. Esteva, A. Kale, R. Paulus, K. Hashimoto, W. Yin, D. Radev, R. Socher, COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization, Npj Digit. Med. 4 (2021). https://doi.org/10.1038/s41746-021-00437-0.

[318] E. Zhang, N. Gupta, R. Tang, X. Han, R. Pradeep, K. Lu, Y. Zhang, R. Nogueira, K. Cho, H. Fang, J. Lin, Covidex: Neural Ranking Models and Keyword Search Infrastructure for the COVID-19 Open Research Dataset, in: 2020: pp. 31–41. https://doi.org/10.18653/v1/2020.sdp-1.5.

[319] A. Köksal, H. Dönmez, R. Özçelik, E. Ozkirimli, A. Özgür, Vapur: A Search Engine to Find Related Protein - Compound Pairs in COVID-19 Literature, in: Proc. 1st Work. NLP COVID-19 (Part 2) EMNLP 2020, Association for Computational Linguistics, Stroudsburg, PA, USA, 2020. https://doi.org/10.18653/v1/2020.nlpcovid19-2.21.

[320] A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, D.C.L. Ngo, Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment, Expert Syst. Appl. 42 (2015) 306–324. https://doi.org/10.1016/j.eswa.2014.08.004.

[321] F. Jin, N. Self, P. Saraf, P. Butler, W. Wang, N. Ramakrishnan, Forex-foreteller: Currency trend modeling using news articles, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2013: pp. 1470–1473. https://doi.org/10.1145/2487575.2487710.

[322] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016: pp. 785–794. https://doi.org/10.1145/2939672.2939785.

[323] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.Y. Liu, LightGBM: A highly efficient gradient boosting decision tree, in: Adv. Neural Inf. Process. Syst., 2017: pp. 3147–3155. https://github.com/Microsoft/LightGBM. (accessed August 16, 2020).

[324] J.C. Wang, T. Hastie, Boosted varying-coefficient regression models for product demand prediction, J. Comput. Graph. Stat. 23 (2014) 361–382. https://doi.org/10.1080/10618600.2013.778777.

[325] H. Qiu, L. Luo, Z. Su, L. Zhou, L. Wang, Y. Chen, Machine learning approaches to predict peak demand days of cardiovascular admissions considering environmental exposure, (n.d.). https://doi.org/10.1186/s12911-020-1101-8.

[326] X. Sun, M. Liu, Z. Sima, A novel cryptocurrency price trend forecasting model based on LightGBM, Financ. Res. Lett. 32 (2020) 101084. https://doi.org/10.1016/j.frl.2018.12.032.

[327] Y. Liang, J. Wu, W. Wang, Y. Cao, B. Zhong, Z. Chen, Z. Li, Product marketing prediction based on XGboost and LightGBM algorithm, in: ACM Int. Conf. Proceeding Ser., 2019: pp. 150–153.

https://doi.org/10.1145/3357254.3357290.

[328] D.M. Blei, Probabilistic topic models, in: Commun. ACM, 2012: pp. 77–84. https://doi.org/10.1145/2133806.2133826.

[329] T. Hofmann, Probabilistic Latent Semantic Analysis, (2013). http://arxiv.org/abs/1301.6705 (accessed March 12, 2022).

[330] D.M. Blei, A.Y. Ng, M.T. Jordan, Latent dirichlet allocation, in: Adv. Neural Inf. Process. Syst., 2002: pp. 993–1022. https://doi.org/10.5555/944919.944937.

[331] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 5228–5235. https://doi.org/10.1073/pnas.0307752101.

[332] L.J. Cao, F.E.H. Tay, Support vector machine with adaptive parameters in financial time series forecasting, IEEE Trans. Neural Networks. 14 (2003) 1506–1518. https://doi.org/10.1109/TNN.2003.820556.

[333] Coronavirus Pandemic (COVID-19) - Statistics and Research - Our World in Data, (n.d.). https://ourworldindata.org/coronavirus (accessed August 16, 2020).

[334] WHO, Coronavirus disease (COVID-19) situation reports in Bnagladesh, World Heal. Organ. (2020) 1. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (accessed August 16, 2020).

[335] Free Coronavirus News Dataset - Updated - AYLIEN, (n.d.). https://blog.aylien.com/free-coronavirus-news-dataset/ (accessed August 16, 2020).

[336] R. Rehurek, gensim: Topic modelling for humans, (2104). https://radimrehurek.com/gensim/index.html (accessed August 16, 2020).

[337] sklearn.metrics.r2_score — scikit-learn 0.23.2 documentation, (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html (accessed August 16, 2020).

[338] sklearn.metrics.mean_absolute_error — scikit-learn 0.23.2 documentation, (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html (accessed August 16, 2020).

[339] sklearn.metrics.mean_squared_error — scikit-learn 0.23.2 documentation, (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html (accessed August 16, 2020).

[340] pandas.Series.mad — pandas 1.2.4 documentation, (n.d.). https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.mad.html (accessed April 18, 2021).

[341] XGBoost Python Package, Python Package Introduction — xgboost 1.4.0-SNAPSHOT documentation, (2020). https://xgboost.readthedocs.io/en/latest/python/python_intro.html (accessed August 16, 2020).

[342] J. Samuel, G.G.M.N. Ali, M.M. Rahman, E. Esawi, Y. Samuel, COVID-19 public sentiment insights and machine learning for tweets classification, Inf. 11 (2020) 314. https://doi.org/10.3390/info11060314.

[343] Swine Flu as Social Media Epidemic; CDC Tweets Calmly – Nielsen, (n.d.). https://www.nielsen.com/us/en/insights/article/2009/swine-flu-as-social-media-epidemic-cdc-tweets-calmly/ (accessed April 3, 2022).

[344] L. Jia, W. Chen, Uncertain SEIAR model for COVID-19 cases in China, Fuzzy Optim. Decis. Mak. 20 (2021) 243–259. https://doi.org/10.1007/s10700-020-09341-w.

[345] A. Mahajan, N.A. Sivadas, R. Solanki, An epidemic model SIPHERD and its application for prediction of the spread of COVID-19 infection in India, Chaos, Solitons and Fractals. 140 (2020) 110156. https://doi.org/10.1016/j.chaos.2020.110156.

[346] I. Korolev, Identification and estimation of the SEIRD epidemic model for COVID-19, J. Econom. 220 (2021) 63–85. https://doi.org/10.1016/j.jeconom.2020.07.038.

[347] H.W. Hethcote, The Mathematics of Infectious Diseases The Mathematics of Infectious Diseases *, Soc. Ind. Appl. Math. 42 (2007) 599–653. https://sci-hub.do/https://epubs.siam.org/doi/abs/10.1137/s0036144500371907 (accessed September 28, 2021).

[348] G. Lee, S.E. Yoon, K. Shin, Simple epidemic models with segmentation can be better than complex ones, PLoS One. 17 (2022) e0262244. https://doi.org/10.1371/journal.pone.0262244.

[349] J.L. Elman, Finding Structure in Time, Cogn. Sci. 14 (1990) 179–211. https://doi.org/10.1207/s15516709cog1402_1.

[350] Ministry of Health Management | Ministry of Health, (n.d.). https://twitter.com/MoHFW_INDIA (accessed November 15, 2021).

List of Publications

Journals

1. Aakansha Gupta, Rahul Katarya, "Social media based surveillance systems for healthcare using machine learning: A systematic review," Journal of Biomedical Informatics, vol. 108. Academic Press Inc., pp. 103500–103500, Jul. 02, 2020, doi: 10.1016/j.jbi.2020.103500. (**Impact Factor: 6.317, Publisher: Elsevier, SCIE**)
2. Aakansha Gupta, Rahul Katarya, "PAN-LDA: A latent Dirichlet allocation based novel feature extraction model for COVID-19 data using machine learning," Comput. Biol. Med., vol. 138, p. 104920, Nov. 2021, doi: 10.1016/j.compbiomed.2021.104920. **(Impact Factor: 4.589, Publisher: Elsevier, SCIE)**
3. Aakansha Gupta, Rahul Katarya, "Deep Embedding for Mental Health Content on Social Media using Vector Space Model with Feature Clusters" Concurr. Comput. Pract. Exp. 34 (2022) e6930. https://doi.org/10.1002/cpe.6930. **(Impact Factor: 1.536, Publisher: Wiley, SCIE)**
4. Aakansha Gupta, Rahul Katarya, "Spatial and temporal distribution of the COVID-19 growth rate in India and its correlation with influencing factors". **(Communicated)**
5. Aakansha Gupta, Rahul Katarya, "Possibility of the COVID-19 third wave in India: mapping from second wave to third wave", Indian J. Phys. Proc. Indian Assoc. Cultiv. Sci. (2022) 1. https://doi.org/10.1007/S12648-022-02425-W. **(Impact Factor: 1.778, Publisher: Springer, SCIE)**
6. Aakansha Gupta, Rahul Katarya, "A Deep-SIQRV Epidemic Model for COVID-19 to Access the Impact of Prevention and Control Measures". **(Communicated)**

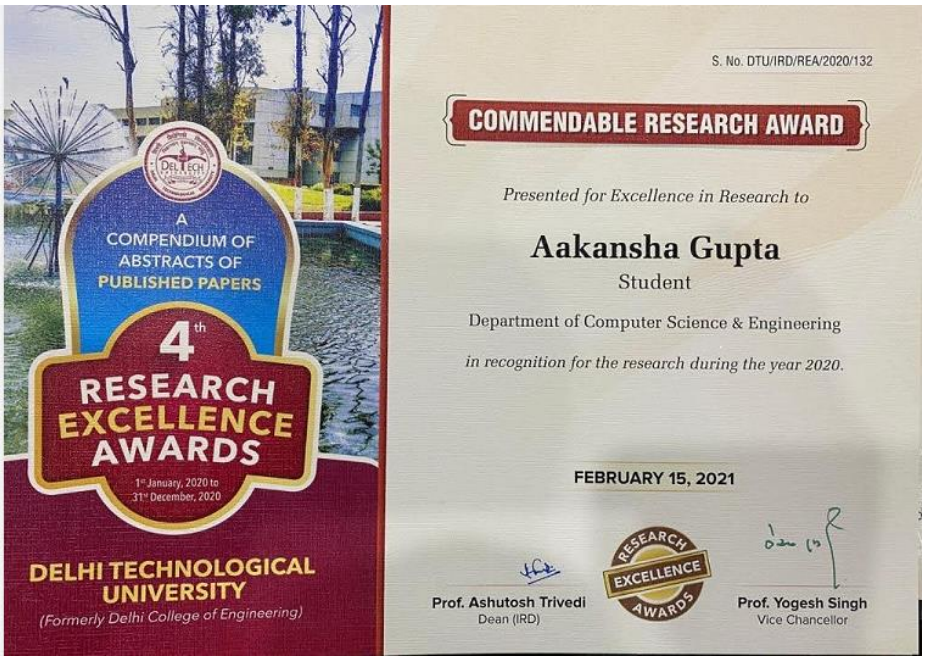
International conferences

1. Aakansha Gupta, Rahul Katarya, "A Novel LDA-based Framework to Forecast COVID-19 Trends". 4th International Conference on Innovative Computing and Communication (ICICC-2021). **Springer**
2. Aakansha Gupta, Rahul Katarya, "Analyzing the Effects of Text Representations on the Performance of Document Clustering in Public Health Tweets" 4th International Conference on Computational Intelligence & Data Engineering (ICCIDE-2021). **Springer**
3. Aakansha Gupta, Rahul Katarya, "Improving document representation using KPCA and clustered word embeddings" 2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT). **IEEE**

4. Aakansha Gupta, Rahul Katarya, "Human mobility based Pandemic Prediction Model" 3rd IEEE International Conference on Advances in Computing, Communication Control and Networking (ICAC3N–21). **IEEE**
5. Aakansha Gupta, Rahul Katarya, "COVID-19 cases prediction based on LSTM and SIR model using social media" 7th International Conference on Data Science and Engineering (ICDSE 2021). **Springer**

Research Award

Commendable Research Award-2021



Commendable Research Award-2022

Biography

(Research Scholar: 2018-22)

Ms. Aakansha Gupta is currently designated as a Senior Research Fellow, a Ph.D. research scholar in the Department of Computer Science, Delhi Technological University, Delhi, India. She has completed her MCA from Jamia Hamdard, Delhi. She has completed an undergraduate degree (B.Sc(H), Computer Science) from the University of Delhi. She has published various research papers in SCIE/SCOPUS/IEEE/SPRINGER indexed International Conferences/Journals. She is UGC-NET qualified with Junior Research Fellowship and later eligible for Senior Research Fellowship. She has more than 3 years of IT industry experience as System Engineer. Her research area of interest includes Artificial Intelligence, Data mining, Machine learning, and Natural Language Processing. She is currently doing her research on public health surveillance using machine learning techniques. She was also awarded the eminent "Commendable Research Award" in 2021 and 2022 from Delhi Technological University, Delhi, India.