Fairness, Engagement, and Discourse Analysis in AI-Driven Social Media and Healthcare

Aditya Singhal

Department of Computer Science
Lakehead University, Canada

AUGUST, 2023

A thesis submitted to Lakehead University in partial fulfillment of the requirements of the degree of

Master of Science in Computer Science

Examination Committee

The thesis of Aditya Singhal, titled Fairness, Engagement, and Discourse Analysis in AI-Driven Social Media and Healthcare, is approved:

.....

Dr. Thiago E Alves de Oliveira
Principal Supervisor, Lakehead University

.....

Dr. Vijay Mago
Co-Supervisor, York University

.....

Dr. Garima Bajwa
Internal Examiner, Lakehead University

.....

Dr. Zahid Butt
External Examiner, University of Waterloo

Declaration

I certify that,

- The work contained in this thesis is original and has been done by myself and under the general supervision of my supervisor(s).
- The work reported herein has not been submitted to any other Institute for any degree or diploma.
- Whenever I have used materials (concepts, ideas, text, expressions, data, graphs, diagrams, theoretical analysis, results, etc.) from other sources, I have given due credit by citing them in the text of the thesis and giving their details in the references. Elaborate sentences used verbatim from published work have been clearly identified and quoted.
- I also affirm that no part of this thesis can be considered plagiarism to the best of my knowledge and understanding and take complete responsibility if any complaint arises.
- I am fully aware that my thesis supervisor(s) are not in a position to check for any possible instance of plagiarism within this submitted work.

Abstract

This thesis addresses the critical concerns of fairness, accountability, transparency, and ethics (FATE) within the context of artificial intelligence (AI) systems applied to social media and healthcare domains. First, a comprehensive survey examines existing research on FATE in AI, specifically focusing on the subdomains of social media and healthcare. The survey evaluates current solutions, highlights their benefits, limitations, and potential challenges, and charts out future research directions. Key findings emphasize the significance of statistical and intersectional fairness in ensuring equitable healthcare access on social media platforms and highlight the pivotal role of transparency in AI systems to foster accountability. Building upon the survey, this thesis delves into an analysis of social media usage by healthcare organizations, with a specific emphasis on engagement and sentiment forecasting during the COVID-19 pandemic. Data collection from Twitter handles of pharmaceutical companies, public health agencies, and the World Health Organization enables extensive analysis. Natural language processing (NLP)-based topic modeling techniques are applied to identify health-related topics, while sentiment forecasting models are employed to gauge public sentiment. The results uncover the impact of COVID-19-related topics on public engagement, highlighting the varying levels of engagement across diverse healthcare organizations. Notably, the World Health Organization exhibits dynamic engagement patterns over time, necessitating adaptable strategies. The thesis further presents latest sentiment forecasting models, such as autoregressive integrated moving average (ARIMA) and seasonal autoregressive integrated moving average with exogenous factors (SARIMAX), which enable organizations to optimize their content strategies for maximum user engagement. Furthermore, discourse analysis is conducted to unravel the factors that shape the content of tweets by healthcare organizations on Twitter. By employing topic modeling and association rule mining techniques, this study uncovers text patterns that significantly influence tweet content across various Twitter accounts. The analysis reveals that establishing a reputable presence on Twit-

ter extends beyond mere tweet popularity, as highly supported association rules do not always translate into increased user engagement. Moreover, the study highlights variations in language use and style among different categories of Twitter accounts. Overall, this thesis makes contributions to the field of NLP for social media and healthcare interventions. By addressing the dimensions of fairness, transparency, and ethics in AI design, it offers insights and practical implications for analyzing public engagement and optimizing content strategies. The integration of AI and NLP techniques empowers healthcare organizations to enhance health literacy, ensure equitable access to healthcare information, and foster maximum public engagement, thereby advancing the field and ultimately improving healthcare outcomes.

Acknowledgements

This thesis work represents the culmination of two years of dedicated effort, and I am grateful for the support of numerous individuals who made this achievement possible. First and foremost, I would like to express my deepest appreciation to my supervisor, Dr. Vijay Mago, for providing me with the opportunity to pursue research under his invaluable guidance. I am grateful for his unwavering support, both morally and financially, and for his guidance throughout my research journey over the past two years.

I am immensely thankful to Lakehead University for granting me access to the resources at the CASES building, which were instrumental in carrying out this work.

I would also like to extend my heartfelt gratitude to my colleagues at DaTALab. The daily discussions, chats, and unwavering support from each of you truly boosted my morale and made the research experience more enjoyable.

Furthermore, I would like to acknowledge the financial support provided by NSERC, which funded the research topic under my supervisor's grant. I am also grateful to the Vector Institute for AI for awarding me the prestigious Vector Scholarship in AI during my graduate studies.

To my family, especially my dear parents and sister, I am profoundly grateful for your unending support and guidance. Despite being thousands of kilometers away from home, you have always stood by my side and encouraged me throughout my master's journey.

Lastly, I want to express my thanks and appreciation to all my friends, family, colleagues, teachers, professors, and well-wishers who have supported me in my educa-

tional pursuits. Your encouragement and belief in my abilities have been instrumental in my success.

I am truly fortunate to have received such tremendous support from everyone, and I am sincerely grateful for each and every individual who has played a part in shaping my educational and research endeavors.

Dedication

This thesis is dedicated in memory of my beloved mother, Mrs. Namita Gupta.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Social media intervention in healthcare has increased in recent years. The advances in natural language processing (NLP) provide a unique opportunity to explore this area. This thesis aims to capture the progress made through an analysis resulting in three seminal research papers.

The main objectives of this thesis are to :

- Present a systematic review of fairness, accountability, transparency, and ethics (FATE) in AI for social media and healthcare.
- Discuss the synergy between public and private healthcare organizations on Twitter and develop an approach to create social media content that maximizes user engagement based on sentiment and engagement analysis using forecasting models.
- Conduct a discourse analysis using association rule mining and causality analysis on Twitter accounts of various healthcare organizations to develop a methodology that can help fine-tune content for the audience.

In Chapter 2 of this thesis, readers will learn about computational methods employed to ensure fairness in artificial intelligence systems used for social media and healthcare domains. With the exponential growth of AI technologies, addressing and mitigating

biases and discrimination perpetuated by these systems is imperative. I study 139 prominent research articles and overall, the objectives of this chapter are to (1) discuss existing definitions of FATE, (2) compare them in terms of computational methods, approaches, and evaluation metrics, and (3) discuss their strengths and drawbacks.

Chapter 3 focuses on the engagement and sentiment of healthcare organizations on Twitter during the COVID-19 pandemic. By examining the content shared by pharmaceutical companies, public health agencies, and non-government organizations (NGOs), this chapter provides insights into the nature of information dissemination and its impact on public engagement. Data were collected from the Twitter handles of 5 pharmaceutical companies, 10 US and Canadian public health agencies, and the World Health Organization (WHO) from January 1, 2017, to December 31, 2021. A total of 181,469 tweets were divided into 2 phases for the analysis, before COVID-19 and during COVID-19, based on the confirmation of the first COVID-19 community transmission case in North America on February 26, 2020. I conducted content analysis to generate health-related topics using NLP-based topic-modeling techniques, analyzed public engagement on Twitter, and performed sentiment forecasting using 16 univariate moving-average and machine learning (ML) models to understand the correlation between public opinion and tweet content.

Chapter 4 uncovers distinctive text patterns that influence tweet content through the application of topic modeling and association rule mining. In this study, I collected a total of 104,347 tweets from January 01, 2020, to December 31, 2022, and the main objectives outlined in this chapter are (1) a discussion on significant text patterns that shape the content of tweets by health agencies and pharmaceutical companies in the US and Canada, and how do they compare with the WHO, and (2) an analysis and evaluation of the impact of word patterns on the content shared by healthcare organizations on Twitter.

To conclude this research, Chapter 5 highlights the main contributions of this thesis and outlines shortcomings and future research directions in this field. The key takeaways of this thesis are:

- **Finding 1:** Statistical and intersectional fairness are highly significant in promoting equitable healthcare practices within social media platforms. Furthermore, transparency in AI systems is critical to ensure accountability and trustworthy decision-making. There is a perpetual need for researchers and practitioners to remain abreast

of the latest advancements in FATE research to navigate the evolving landscape effectively.

- **Finding 2:** People engage more on topics such as COVID-19 than medical trials and customer experience on Twitter. In addition, there are notable differences in user engagement levels across organizations. Global organizations, such as WHO, show wide variations in engagement levels over time. The sentiment forecasting method discussed presents a way for organizations to structure their future content to ensure maximum user engagement.

- **Finding 3:** NLP methods, such as topic modeling, help identify the overall themes and topics of the tweets, but association rule mining can help identify which words, phrases, or language patterns are associated with higher or lower tweet popularity, allowing organizations to adjust their messaging and communication strategies accordingly. Using popular association rules also significantly increases the probability of a tweet getting reshared across all categories.

The research conducted during this work is open-sourced and readily available in GitHub repositories[1][2].

Overall, this thesis addresses the critical aspects of FATE in AI, explores the engagement and sentiment of healthcare organizations on social media, and provides a comprehensive analysis of discourse and language patterns in their tweets. By examining these interconnected topics, the thesis aims to offer valuable insights to researchers, practitioners, and policymakers in the domains of AI, healthcare, and social media. Through its rigorous examination and analysis, this thesis contributes to the ongoing discourse surrounding FATE in AI and its implications for social media and healthcare.

[1]https://github.com/manmeetkaurbaxi/Sentiment-Forecasting-on-tweets
[2]https://github.com/aditya-ml/Association-Rule-Mining

Chapter 2

Towards FATE in AI for Social Media and Healthcare: A Systematic Review

All of this chapter is submitted at a reputed journal as [178]:
- Singhal, A., Tanveer, H., & Mago, V. (2023). Towards FATE in AI for Social Media and Healthcare: A Systematic Review

Over the course of my degree, I researched topics related to fairness, accountability, transparency, and ethics to expand my knowledge in the field. As a result, I performed a systematic review to understand and analyze different research techniques and research applications of FATE in social media and healthcare.

Keywords : fairness, accountability, transparency, ethics, artificial intelligence, social media, healthcare

2.1 Introduction

2.1.1 Background

Machine learning algorithms are utilized by all stakeholders in today's world. Most fields, from governance to financial decision-making, and medical diagnosis to security assessment, depend on artificial intelligence (AI) to deliver results. On the surface, this progression towards automation has clear benefits: it is fast and reliable, and cost-effective for businesses over time [131]. However, as research in AI continues to advance at a rapid pace, it is becoming increasingly important to ensure that its development and deployment are guided by principles of fairness, accountability, transparency, and ethics.

The vast amount of user data available on social media platforms (SMPs) can be used to identify patterns, trends, and behaviour. SMPs such as Twitter are predominantly used by young and urban residents [127]. They also have minimum age requirements, leading any machine learning algorithm trained on data from these sources to be biased toward a certain demographic. The wide availability of social media data is also opportunistic for health-related research [110]. However, lack of ethical oversight at the data collection stage of a research project could lead to the inclusion of data from users who did not consent to it, thereby raising questions on *'who is a participant of the study?'*. The content on SMPs is also heavily influenced by *'social'* processes and can not be taken at its face value. Certain topics might generate traction from users of particular areas or demographics [177], and the trustworthiness of data continues to be a challenge [110]. The lack of availability of code behind machine learning algorithms in propriety software makes it difficult to analyze the underlying patterns which may cause discriminative decisions.

Misinformation refers to the inadvertent dissemination of false information on a topic, while *disinformation* is the intentional spread of false information for motives such as financial gain, fame, or damaging the reputation of others [100]. The proliferation of false information is prevalent on social media, and during the COVID-19 pandemic, there was widespread misinformation about vaccines, including unfounded claims that they were harmful. Such misinformation led to doubts about the government and vaccine hesitancy, posing risks to public health and efforts to control the spread of COVID-19. To counter this issue, AI is being employed to help identify and label reliable and high-quality information for users. Reliable AI systems are trained using accurate health information

from reputable sources such as non-profit health organizations or government agencies, ensuring that the information provided is based on sound scientific evidence. In addition, corporations are advised to prioritize transparency to build trust among the public and foster a better community. Transparent practices can enhance accountability and credibility, leading to increased trust from the people and promoting a positive environment of mutual trust and understanding.

On the positive side, social media can serve as a platform for users to share new health information, allowing the health sector to potentially access more medical insights and knowledge [158]. However, the drawbacks of social media, such as the lack of verifiability and potential misinformation, need to be carefully addressed to ensure accurate and reliable health information dissemination. The FATE research focuses on evaluating the fairness and transparency of AI models, developing metrics to assess the accountability of AI systems, and designing frameworks for responsible and ethical AI development. The ethical implications of using algorithms and AI are closely dependent on the practices of transparency and accountability [207]. Algorithmic systems can be viewed as socio-technical systems that are involved in many different sectors, such as culture, programming, laws and more. One way to avoid discrimination is to have a human in the loop of these algorithmic processes when needed. For example, when the US judicial system COMPAS decides on the likelihood of a prisoner committing another crime after leaving prison, a judge should review the decision of the AI first in order to check the accuracy of the decision. Since the current AI systems are expected to have some bias in their decision models, ensuring that researchers are implementing these models in an organized, ethical, and systematic manner, would make it easier to enforce accountability of actions [84]. Efforts are being made by computer scientists to make AI more transparent by revealing the decision-making process that leads to the final AI answer [91]. This helps in identifying problems or biases and holding individuals accountable in case of failures. The European Union has recommended seven key principles to ensure ethical AI, including human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, nondiscrimination and fairness, social and environmental well-being, and accountability.

Table 2.1: An overview of existing survey articles focusing on FATE. *A = Definitions, B = Computational methods and approaches, C = Evaluation metrics*

Paper	Fairness			Accountability			Transparency			Ethics		
	A	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
[130]	✓	✓	✓									
[68]											✓	✓
[16]		✓	✓	✓								
[12]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
[207]				✓	✓	✓						
[2]								✓	✓			
[23]							✓					
[28]	✓					✓	✓	✓	✓			✓
[30]	✓			✓			✓		✓			
[76]										✓		✓
[92]						✓			✓	✓	✓	✓
[148]						✓			✓			
Our paper	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

2.1.2 Motivation

Studies on the FATE of AI in social media surveillance have focused on ensuring that AI systems used for monitoring online content and activity do not perpetuate existing biases or discrimination. For example, research has shown that algorithms used in social media surveillance may have biases against certain groups, or that algorithms used to detect hate speech may not be effective in detecting it against all groups. Recent research has also mainly focused on one aspect of understanding machine learning models. The research in AI ethics is heavily influenced by geographic locations and socio-economic factors [76]. There have been several discussions on the best practices for evaluating work produced by explanatory AI (XAI) and gap analyses performed on model interpretability in AI [67], [30]. The latest developments in machine learning interpretability have also been reviewed previously [28]. Table 4.1 provides an overview of existing review studies discussing FATE in different forms. Therefore, the motivation behind this survey is to present a comprehensive overview of the various computational methods for FATE and provide direction for future research in the field.

We aim to address the following research questions in this work:

Table 2.2: Search strategy for finding research articles. *T = Search term, G = Group, Quality = {fairness, accountability, transparency, ethics}*

	G1	**G2**	G3
T1	Quality	Natural language processing	Social media
T2		Artificial intelligence	Healthcare
T3		Computer science	

RQ1: What are the existing solutions to FATE (Fairness, Accountability, Transparency, and Ethics) when discussing healthcare on Social Media Platforms (SMPs)?

RQ2: How do the different solutions identified in response to RQ1 compare to each other in terms of computational methods, approaches, and evaluation metrics?

RQ3: What is the strength of evidence supporting the different solutions?

The objective of this research is to identify gaps in the current literature and complement existing work in the FATE space by showcasing how various techniques, themes, and contextual considerations can be combined to support social media interventions in healthcare settings.

2.1.3 Research Methodology

Our research methodology is based on the approach presented by the authors of [102]. We utilized Google Scholar[1], the largest repository of scholarly articles, to perform a strategic search using Table 4.2 as a filter to identify research papers relevant to our study. Each of the groups in the table can be customized to retrieve different sets of literature, with the aim of finding the intersection of these sets. This search strategy involves using the AND and OR operators, where the OR operator can be used within the groups and the AND operator between the groups.

The search strategy employed for this study can be summarized as follows: (T1G1 AND T1G2) AND (T1G1 AND T2G2) AND (T1G1 AND T3G2) OR (T1G1 AND T1G3) AND (T1G1 AND T2G3). Initially, this search yielded a substantial number of results, which were then filtered using the following steps: (1) considering articles published

[1]https://scholar.google.com

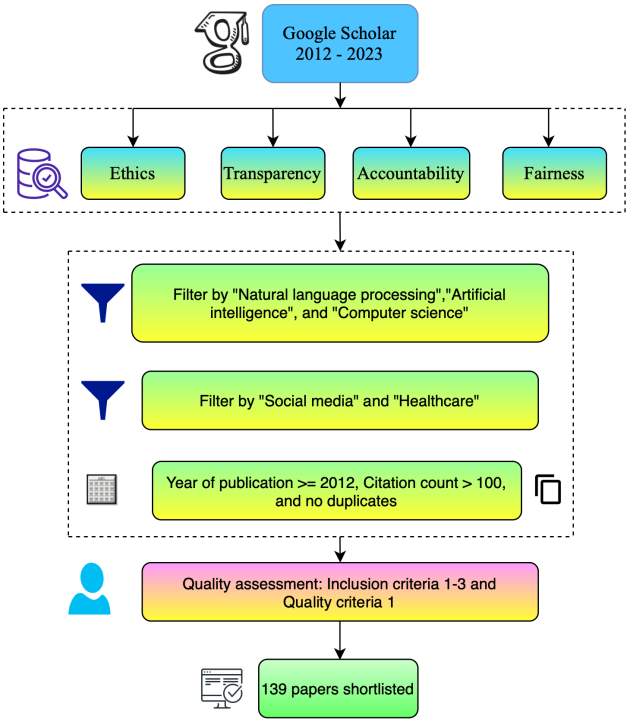


Figure 2.1: Overview of the search strategy and research methodology.

after 2012, (2) including articles with a generally high citation score (>100), with some exceptions for recent articles with citations < 100, and (3) removing all duplicate articles. Subsequently, a quality assessment of all articles was conducted based on inclusion (IC) and quality criteria (QC). The inclusion criteria consisted of IC1 (the study's main concern is FATE while discussing healthcare on SMPs), IC2 (the study is a primary study presenting empirical results), and IC3 (the study focuses on definitions, computational methods, approaches, and evaluation metrics). The quality criteria included QC1 (clear statement of the research aim). These criteria were applied through a three-stage process: abstract inclusion criteria screening, full-text inclusion criteria screening, and full-text quality screening. This process helped us determine the relevance and quality of the articles for our study.

Figure 4.1 outlines the structure and overall, this survey provides an in-depth understanding of one of the most socially important problems in AI for new researchers. The article is structured as follows: Sections 2.2, 2.3, 2.4, and 2.5 cover the definitions, computational methods, approaches, and evaluation metrics for FATE in AI. Section 2.6 provides an overview of FATE in datasets, while section 2.7 offers a discussion on the topic. Finally,

future research directions and conclusions are presented in sections 2.8 and 2.9, respectively.

2.2 An Overview of Fairness

2.2.1 Definitions, Computational Methods, and Approaches to Fairness

The extent to which the general public understands the definition of fairness varies [169]. There are several different definitions that have been proposed in the context of artificial intelligence.

Calibrated fairness [172]

It refers to the balance between providing equal opportunities for all individuals and accommodating for their differences and needs. For example, in social media, a *calibrated fair* algorithm could ensure that all users have equal access to opportunities, such as visibility, while also taking into consideration specific factors, such as language or location, to provide a personalized experience. In healthcare, a *calibrated fair* algorithm could ensure that all patients have access to the same standard of care while accounting for their age and health status to provide the best possible treatment plan. The goal is to strike a balance between treating everyone the same and taking into account individual differences to provide the most equitable outcomes. Fairness metrics, such as the True Positive Rate Difference (TPRD) [130], False Positive Rate Difference (FPRD) [212], and Equal Opportunity Difference (EOD) [152] can be used to assess the level of calibrated fairness. Other commonly used computational methods to achieve calibrated fairness are:

1. Pre-processing: transforming the original data set to remove or reduce the effect of sensitive attributes (such as race and gender) on the outcome of a machine learning model [210].

2. In-processing: incorporating fairness constraints into the training process of the model to ensure that the model is calibrated with respect to the sensitive attributes [210].

3. Post-processing: adjusting the output of a model after it has been trained to ensure that it is calibrated with respect to the sensitive attributes [210].

4. Adversarial training: training the machine learning model on adversarial examples, or examples that are specifically designed to challenge the model's ability to make fair predictions [191].

Statistical fairness

It takes into account various factors, such as demographic information, that may be relevant to the notion of fairness in a specific context. Some commonly used statistical definitions of fairness include demographic parity, equal opportunity, and equal treatment [35]. The *'demographic parity'* measure can be utilized to minimize data bias by augmenting matrix-factorization objectives with penalty functions [213], while the *'equal opportunity'* metric is important to ensure decisions are free from bias [218]. In the context of social media, individual definitions of fairness might include issues such as unbiased content moderation, fair representation of diverse perspectives and voices, and transparency in the algorithms used to curate and rank content. Commonly used computational metrics are:

1. Equalized odds: measures fairness by comparing the true positive rate and false positive rate for different groups [63].

2. Theorem of equal treatment: measures fairness by comparing the treatment of similar individuals belonging to different groups [124].

Intersectional fairness [64]

This metric takes into account multiple and intersecting aspects of identity, such as race, gender, and socio-economic status, when making decisions about people. The goal is to

ensure that people are not discriminated against, as these intersections can compound and result in greater marginalization and unequal treatment. In the context of social media, an algorithm that takes into account intersectional fairness would ensure that content is not recommended or censored in a biased manner based on a user's race, gender, and socio-economic status, while in the context of healthcare, an algorithm that considers intersectional fairness would ensure that medical treatments and resources are not disproportionately allocated. The best approach to implementing intersectional fairness is through the *worst-case disparity* method. This entails assessing each subgroup individually and comparing the best and worst outcomes to determine the accuracy of the fairness score. The ratio of the maximum and minimum scores is then calculated, and the closer the ratio is to 1, the fairer the outcome is [64]. Other commonly used methods include:

1. Constraints-based methods: the algorithm is designed to respect certain fairness constraints, such as equal treatment for different groups based on multiple attributes through mathematical optimization [214].
2. Causal inference methods: ensure that the algorithm's outputs are not biased by considering the causal relationships between the inputs and outputs [29].
3. Decision trees and rule-based systems: to ensure that the algorithm's decisions are based on appropriate factors and are not biased [166].

The supervised ranking, unsupervised regression, and reinforcement aspects of fairness evaluation can be done using *pairwise evaluation* [141]. This involves evaluating the performance of an AI model by comparing its output to a set of predefined pairs of input data.

2.3 An Overview of Accountability

2.3.1 Definitions, Computational Methods, and Approaches to Accountability

It refers to the notion that individuals or organizations using AI should be responsible and answerable for the consequences of their systems. This includes the responsibility to

Table 2.3: Fairness evaluation metrics with mathematical formulation. *FP = False Positive, FN = False Negative, TP = True Positive, TN = True Negative, A = binary attribute representing a demographic group*

Metric	Formula	Description
Equal Opportunity [35]	$P(\frac{FP}{y=1}) = P(\frac{FP}{A=1,y=1}) - P(\frac{FP}{A=0,y=1})$ where y is the true label	An AI model's positive outcomes are not systematically skewed towards or against certain groups of people.
Equal Odds [35]	$P(\frac{y=1}{p>t,y=1}) = P(\frac{y=1}{p>t,y=0}) = P(\frac{y=0}{p<=t,y=1}) = P(\frac{y=0}{p<=t,y=0})$ where y is the true label, p is the predicted probability of positive class, and t is a threshold.	The false positive rate and the false negative rate are equal across different groups of people.
Demographic Parity [35]	$P(\frac{y=1}{A=1}) = P(\frac{y=1}{A=0})$ where y is the predicted label.	The proportion of positive outcomes for different groups of people is equal.
Statistical Parity [81]	$P(\frac{Y=1}{A=a}) = P(Y=1)$ for all a in A	The proportion of favorable outcomes is the same for all groups.
Accuracy [82]	$\frac{TP+TN}{TotalPopulation}$	The proportion of all predictions that are correct.
False Positive Rate (FPR) [212]	$\frac{FP}{FP+TN}$	The proportion of negative instances that are incorrectly classified as positive.
False Negative Rate (FNR) [145]	$\frac{FN}{FN+TP}$	The proportion of positive instances that are incorrectly classified as negative.
True Positive Rate (TPR) [130]	$\frac{TP}{TP+FN}$	The proportion of positive instances that are correctly classified as positive. Also known as sensitivity or recall.
True Negative Rate (TNR) [145]	$\frac{TN}{TN+FP}$	The proportion of negative instances that are correctly classified as negative. Also known as specificity.
Positive Predictive Value (PPV) [209]	$\frac{TP}{TP+FP}$	The proportion of instances that are predicted as positive that are actually positive.
Negative Predictive Value (NPV) [209]	$\frac{TN}{TN+FN}$	The proportion of instances that are predicted as negative that are actually negative.
False Discovery Rate (FDR) [74]	$\frac{FP}{FP+TP}$	The proportion of instances that are predicted as positive that are actually negative.
False Omission Rate (FOR) [126]	$\frac{FN}{FN+TN}$	The proportion of instances that are predicted as negative that are actually positive.
Positive Likelihood Ratio (LR+) [199]	$\frac{TP}{FP}$	Indicates how much more likely a positive result is to occur when the condition is present than when it is absent.
Negative Likelihood Ratio (LR-) [199]	$\frac{FN}{TN}$	Indicates how much more likely a negative result is to occur when the condition is absent than when it is present.

ensure that the system operates in an ethical manner, with the goal of providing equitable and accurate outcomes. There are several definitions of accountability in AI, including:

Legal accountability [216]

It refers to the legal obligations of entities involved in the design, development, deployment, and use of AI systems for healthcare purposes in social media. This includes the responsibility for ensuring that the AI systems are developed and used in accordance with applicable laws and regulations, as well as the responsibility for any negative consequences or impacts that may result from their use. Legal accountability may also extend to issues such as data protection and privacy, and the responsibility for ensuring that

AI systems are not used for discriminatory or unethical purposes. The commonly used conceptual computational methods are:

1. Transparency: Ensuring that AI systems are transparent and that their decision-making processes can be explained and understood [22].
2. Documentation: Keeping records of systems' design, development, and testing processes, as well as the data used to train them [52].
3. Auditing: Conducting independent assessments of AI performance and accuracy to verify compliance with legal requirements [153].
4. Regulation: Implementing legal frameworks that establish standards and requirements for the development, deployment, and use of AI systems [153].
5. Adjudication: Establishing procedures for resolving disputes and grievances related to the use of AI systems [99].

Ethical accountability [137]

It involves ensuring that the decisions made by AI systems are transparent, justifiable, and in line with the values of society. This includes issues such as data privacy, informed consent, and ensuring that AI systems do not perpetuate existing biases and discrimination. The ethical considerations around the use of AI in healthcare include topics such as the protection of patient privacy, the use of sensitive health data, and the potential for AI systems to reinforce existing health disparities [94]. Stakeholders involved in the development and deployment of AI in healthcare have a responsibility to ensure that ethical principles are integrated into the design and implementation of these systems, and that the outcomes of their use are regularly monitored and evaluated for any ethical concerns. Some common methods include:

1. Ethical Impact Assessment: involves identifying the ethical risks and benefits of the system, and determining the trade-offs between them [208].
2. Value Alignment: invloves incorporating ethical principles and values into the design and development of the system, and ensuring that its behavior is consistent with these values. [9].

3. Transparency and Explanation: achieved by providing clear and concise explanations of how the system works, and by making its data and algorithms open and accessible [85].

4. Stakeholder Engagement: involves engaging stakeholders, including users, developers, and experts, in the development and evaluation of AI systems [196].

Technical accountability [201]

It refers to the responsibility of the developers and designers of AI systems to ensure that their technology meets certain standards of functionality, security, and privacy. This includes having appropriate systems in place to monitor and manage the AI algorithms, as well as addressing any technical issues that arise. In the context of social media and healthcare, technical accountability also involves considering how AI technologies can be used to support ethical decision-making, such as ensuring that user privacy is protected and that decisions are made in a fair and transparent manner [149]. Some of the commonly used methods:

1. Logging: logging all inputs, outputs, and decisions can be used to track the system's performance and identify potential issues [101].

2. Auditing: to assess their performance, identify potential biases, and ensure that they are aligned with ethical and legal standards [161].

3. Transparency, Model interpretability, and Explainability: can be designed to provide users with clear explanations of their decision-making processes, which can help to increase trust in the system and reduce the risk of ethical and legal violations [201].

Societal accountability [200]

It refers to the responsibility of the stakeholders to ensure that the use of AI systems aligns with the values and interests of society as a whole. This includes issues such as privacy, transparency, and fairness, as well as broader social, cultural, and economic factors that can be affected by AI systems. To ensure societal accountability, it may be necessary for

stakeholders to engage in public consultation, to develop regulations and standards that ensure that AI systems are used ethically and transparently, and to promote transparency and public understanding of how AI systems work and what they are being used for. Ultimately, it means that the development and use of AI systems is guided by the principles of responsible innovation, and that the interests of society are taken into account in all stages of their lifecycle. Other methods are:

1. Regulation and standardization: development of regulations and standards for the design and use of AI systems. This can help ensure that AI systems are accountable to society and that they operate in a way that protects the rights and interests of all stakeholders [97].

2. Public-private partnerships: collaboration between government agencies, private companies, and civil society organizations to ensure that AI systems are developed and used in a way that is accountable to society [163].

Accountability can be ensured by implementing transparency and fairness into the algorithms, designing systems with privacy in mind, and conducting regular audits and evaluations to assess the performance of the AI system. Researchers have proposed a method for holding companies accountable for their actions related to AI [207]. They argue that it is crucial to first identify the specific decision-makers within the company who are responsible for the error in question. This is essential for ensuring fair judgment. The person or group responsible for determining accountability should be well-versed in the various legal, political, administrative, professional, and social perspectives related to the topic of the error to ensure that the judgement is fair and unbiased. Finally, the consequences for the decision-makers should be tailored to the specific areas of their responsibility, and the level of responsibility of each individual decision-maker within the company's hierarchy should be considered when determining them. Algorithms such as decision trees and regression models are more interpretable than others [179]. With the widespread adoption of deep learning methods in decision models, explainability in AI (XAI) also presents a way to interpret the model by humans.

Table 2.4: Accountability evaluation metrics with mathematical formulation. Accuracy, False Positive Rate, and False Negative Rate metrics are also suitable for evaluating accountability in AI systems, as discussed in Table 4.1. *FP = False Positive, FN = False Negative, TP = True Positive, TN = True Negative*

Metric	Formula	Description
Fairness [106]	$\frac{\#TP\_for\_groupA}{\#actual\_positives\_for\_A} - \frac{\#TP\_for\_groupB}{\#actual\_positive\_for\_B}$	Measures whether the AI system treats different groups fairly
Explainability [95]	$\frac{\#explanations\_provided}{\#decisions\_made}$	Measures how well the AI system can explain its decision-making process
Consistency [26]	$1 - \frac{\#changes\_to\_output}{\#decisions\_made}$	Measures how consistent the AI system's outputs are over time
Robustness [205]	$\frac{\#correctoutputs}{\#decisionsmade}$	Measures how well the AI system performs under unexpected conditions
Precision [105]	$\frac{TP}{TP+FP}$	The proportion of true positive predictions among all positive predictions.
Recall (Sensitivity) [105]	$\frac{TP}{TP+FN}$	The proportion of true positive predictions among all actual positive instances.
Specificity [86]	$\frac{TN}{TN+FP}$	The proportion of true negative predictions among all actual negative instances.
F1 Score [72]	$2 * \frac{Precision*Recall}{Precision+Recall}$	The harmonic mean of precision and recall.
Confusion Matrix [126]	$[[TP, FP], [FN, TN]]$	A table used to evaluate the performance of a classifier.
Pandora [147]	Five-fold cross validation for cluster prediction accuracy	It is a hybrid of human and system-generated observations to explain system failure for analysis and debugging.

2.4 An Overview of Transparency

2.4.1 Definitions, Computational Methods, and Approaches to Transparency

Transparency in AI refers to the degree to which the internal workings of an AI system can be understood by humans [2]. It involves providing explanations for how the system makes decisions, understanding the data that was used to train the system, and ensuring that the system is not biased or discriminatory. The issues of transparency and privacy are often at cross-heads. For example, while analyzing mental health data on social media platforms, the challenge is not with the identifying attributes of individual users (as the data is often aggregated), but with how that data is utilized [37]. There are several different definitions of transparency depending on the specific context and use case:

Algorithmic transparency [47]

It refers to the ability to understand how an AI algorithm or model arrives at its outputs or decisions. In the context of social media for healthcare, transparency can be defined as the ability to clearly understand the processes and methods used to create, disseminate, and evaluate social media interventions for healthcare purposes [186]. This includes being able to understand the data sources used to inform the interventions, the algorithms or models used to analyze the data and create the interventions, and the criteria used to evaluate the effectiveness of the interventions. Transparency is important because it allows for the identification and mitigation of potential biases or errors in the interventions, and helps to build trust with stakeholders, including patients, healthcare providers, and regulators. There are several computational methods that can be used to increase algorithmic transparency, such as:

1. Feature importance analysis: involves identifying the features or variables that have the most significant impact on the model's output. By doing so, it helps to understand the model's decision-making process [197].
2. Model interpretability: involves designing models in such a way that their output can be easily understood and interpreted by humans. For example, decision trees are considered interpretable because their output can be visualized as a series of decision nodes [117].
3. Explanation generation: involves generating explanations for the model's output. These explanations can be in the form of natural language or visualizations, and they help to provide insight into the model's decision-making process [187].

Data transparency

It refers to the ability to understand how data is collected, stored, and used in the development of an AI system [19]. In the context of healthcare, data transparency refers to the extent to which healthcare organizations and providers are open and clear about the collection, storage, and use of patient data in the design and implementation of their social media campaigns [80]. This includes providing patients with clear information about what data is being collected, how it will be used, who will have access to it, and how

it will be protected. By being transparent about data collection and use, healthcare organizations can build trust with patients and promote more active engagement in social media-based health interventions. This can ultimately lead to better health outcomes for patients, as they are more likely to participate in interventions that they feel comfortable with and have confidence in. Computational methods for data transparency can include:

1. Data visualization: involves creating graphical representations of data in order to make it easier for users to understand and interpret [113].
2. Data profiling: involves analyzing data to understand its structure, quality, and content, which can help identify issues such as missing values or inconsistencies [13].
3. Data lineage analysis: involves tracing the movement of data through various systems and processes to ensure its accuracy and reliability [23].

Process transparency

It refers to the ability to understand the steps taken to develop and deploy an AI system, including the testing and validation processes used [111]. In the context of social media and healthcare, it refers to the transparency of the decision-making process that determines which health-related information is prioritized, displayed, and disseminated on social media platforms [148]. This can include transparency around the algorithms and other computational methods used to curate and display health-related content, as well as the policies and guidelines used to moderate user-generated content related to health. By increasing process transparency, users can have more confidence in the information and interventions being presented to them, and researchers can have greater trust in the data they are analyzing. There are several computational methods that can be used to increase process transparency in AI systems:

1. Data provenance tracking: involves tracking the origin, processing history, and movement of data throughout the AI system. This helps to ensure that the data used in the system is reliable and can be traced back to its source [27].
2. Model interpretability: involves developing algorithms and tools that can help explain how an AI system makes decisions. Techniques such as feature importance

analysis [197], decision trees [55], and partial dependence plots [219] can help to uncover how the model arrives at its predictions.

3. Explanation generation: involves generating natural language or visual explanations for the decisions made by an AI system. Techniques such as saliency maps, LIME (Local Interpretable Model-agnostic Explanations) [215], and SHAP (SHapley Additive exPlanations) [120] can help to generate these explanations.

4. Auditability and monitoring: involves building auditing and monitoring capabilities into the AI system. This can include monitoring the system's performance, detecting bias or other ethical issues, and identifying when the system is not performing as intended [175].

5. Open-source development: involves developing AI systems in an open and transparent manner, where the code, data, and models are publicly available. This allows for greater scrutiny and accountability of the system by external stakeholders, such as regulators or the general public [25].

Explainability

It refers to the ability to provide a clear and understandable explanation of how an AI system arrived at a particular decision or recommendation [89]. In the context of social media intervention for healthcare, explainability can involve understanding how an AI system is processing social media data, how it is identifying relevant information, and how it is making recommendations or decisions based on that data [150]. It can also involve understanding the factors that influenced the system's decision-making, such as the data used to train the model or the specific features that were weighted more heavily in the decision process. To achieve explainability in AI systems for social media intervention in healthcare, various methods can be used, including techniques for feature selection, model interpretability, and visualizations. These methods can help healthcare professionals to better understand the underlying mechanisms of an AI system and the factors that contribute to its decision-making process. Some common methods include:

1. Decision trees: Decision trees are graphical representations of the decision-making process of a model. They can be used to explain how the model is making decisions and which factors are most influential [55].

2. LIME (Local Interpretable Model-Agnostic Explanations): LIME is a method for explaining the predictions of any machine learning model. It works by generating a simpler, interpretable model that approximates the behavior of the original model [215].

3. SHAP (SHapley Additive exPlanations): SHAP is a method for explaining the output of any machine learning model. It works by computing the contribution of each input feature to the final prediction [120].

4. Counterfactual explanations: Counterfactual explanations involve identifying the minimal set of changes to the input features that would result in a different output from the model. They can be used to explain why a certain prediction was made and what could have been done differently to change the outcome [181].

Interpretability

It refers to the ability to understand the meaning and implications of the decisions made by an AI system, including how they impact different groups of people [28]. Interpretability also refers to the ability of an AI system to provide a clear and understandable explanation for its decisions or recommendations to healthcare professionals, patients, and other stakeholders [8]. This is particularly important in healthcare, where the consequences of AI decisions can be critical to patient outcomes. An interpretable system enables stakeholders to understand how the AI arrived at its recommendation and can help build trust in the system. Interpretability techniques in AI involve designing models with clear and understandable features, such as decision trees or rule-based systems [10]. These methods can help identify the factors that influenced the AI's decision, making it easier to understand and explain the outcome. Other techniques include generating visualizations, such as heatmaps or saliency maps, which highlight the areas of an input that had the most significant impact on the model's output. By providing clear explanations of the model's decision-making process, these techniques can help stakeholders better understand and trust the AI system. Some of the computational methods are:

1. Partial Dependence Plot (PDP): PDP shows the relationship between the target variable and one or two input variables while controlling for the effects of other input variables. This shows how the model is making predictions and how the input variables are affecting the output [219].

Table 2.5: Transparency evaluation metrics with mathematical formulation

Metric	Formula	Description
Completeness [203]	$\frac{\#available\_data\_points}{\#total\_data\_points}$	The extent to which all relevant information is available
Timeliness [40]	$\frac{\#data\_points\_available\_within\_timeframe}{\#required\_data\_points}$	The extent to which data is available in a timely manner
Relevance [217]	$\frac{\#relevant\_data\_points}{\#data\_points}$	The extent to which data is applicable to the problem at hand
Accessibility [204]	$\frac{\#data\_points\_that\_can\_be\_obtained}{\#data\_points}$	The extent to which data is easy to obtain and use
Data Provenance [27]	$\frac{\%data\_with\_known\_source}{chain\_of\_custody}$	Involves tracking the origin, processing history, and movement of data throughout the AI system.

2. Local Interpretable Model-Agnostic Explanations (LIME): LIME is a post-hoc method that explains the output of any classifier by approximating it with an interpretable model locally. This shows how the model is making decisions for a specific instance [215].

3. Model Distillation: It is the process of training a simpler model that approximates the decision boundaries of a more complex model. This can help in creating a simpler and more interpretable model that still maintains the performance of the original model [139].

Overall, transparency in AI is important for ensuring accountability, fairness, and ethical use of AI systems. It helps build trust with users and stakeholders, and can also help identify and address biases or errors in the system.

2.5 An Overview of Ethics

2.5.1 Definitions, Computational Methods, and Approaches to Ethics

In AI, ethics refers to the study and practice of developing and implementing AI technologies in a manner that is fair, transparent, and beneficial to all stakeholders [109]. The goal of ethical AI is to ensure that AI systems and their decisions are aligned with human values, respect fundamental human rights, and do not result in harm or discrimination against individuals or groups. This includes considerations of privacy, data protection, bias, accountability, and explainability [107]. In the context of social media, dig-

ital surveillance of public health data from social media platforms should be guided by the principles of 1) beneficence: surveillance must lead to improvement in public health outcomes; 2) non-maleficence: use of data should not erode public trust; 3) autonomy: informed consent of users or anonymizing of identifying details; 4) equity: equal opportunities to individuals for public health interventions, and 5) efficiency: building legal mandates to ensure continuous access to web platforms and decision-making algorithms [6]. AI-mediated healthcare treatments must account for affordability and equality among the masses, and while nascent, health tech in the field of *patient-centric* models is no longer science fiction, wherein scientifically tailored medicines are prescribed to the patients [70]. There are many different definitions of ethics, depending on the context in which the term is used:

Philosophical ethics

Refers to the concept of ensuring that AI systems are designed and used in ways that respect human autonomy, dignity, and privacy [96]. In the context of social media intervention for healthcare, philosophical ethics in AI refers to the study and application of ethical principles and values to the development and use of AI-powered tools and technologies for healthcare interventions via social media [143]. It involves examining the potential benefits and risks of using AI to collect, analyze, and interpret health-related data from social media platforms, as well as ensuring that the use of such technologies aligns with ethical principles such as respect for privacy, autonomy, beneficence, and non-maleficence. It also involves considering the potential biases that may arise in the development and use of these technologies, and taking steps to mitigate these biases to ensure that the use of AI in social media intervention for healthcare is fair, just, and equitable for all individuals involved. Ultimately, the aim is to promote the development and use of AI technologies that improve health outcomes, while minimizing the potential risks and harms that may arise from their use. Some examples of computational methods for philosophical ethics include:

1. Simulation and modeling: These are techniques that allow ethical dilemmas to be simulated and modeled, providing insights into the likely outcomes of different ethical decisions [41].

2. Game theory: This is a mathematical framework that can be used to model and analyze decision-making in social situations, including ethical dilemmas [171].

3. Data analytics: This involves the use of statistical methods and machine learning algorithms to analyze data and identify patterns or insights related to ethical questions or dilemmas [182].

Moral ethics

Refers to the ethical considerations that need to be taken into account while using social media for healthcare interventions [75]. This includes ensuring that the privacy and confidentiality of patient data are maintained, that the patient's autonomy and consent are respected, and that the use of social media platforms does not result in any harm to the patient [20]. It also involves ensuring that the interventions are based on evidence-based practices and that the potential benefits of the intervention outweigh the potential risks. Finally, it involves being transparent about the use of social media for healthcare interventions and communicating the risks and benefits to all stakeholders involved [184]. Some of the computational methods are:

1. Data visualization tools: These tools can be used to present complex ethical data in a clear and accessible way, making it easier for healthcare professionals and other stakeholders to understand and make informed decisions [44].

2. Sentiment analysis: Language and sentiment of social media posts related to healthcare interventions can help identify any ethical issues or concerns that may arise, such as biases or stigmatization of certain patient groups [118].

3. Crowdsourcing platforms: developed to gather feedback from a diverse group of individuals on the ethical implications of the AI system and its recommendations. This can help ensure that the system takes into account a range of perspectives and values, and can identify potential ethical concerns that may have been overlooked by the development team [87].

Professional ethics

In the context of healthcare and social media intervention, professional ethics involves a set of guidelines and principles that guide the behavior of HCPs who are using social media as part of their practice [198]. This might include guidelines around patient privacy, confidentiality, informed consent, and the appropriate use of social media platforms (i.e., avoiding conflicts of interest or biased behavior) for sharing health information. Computational methods for enforcing professional ethics might include:

1. Automated systems: for monitoring healthcare professionals' behavior on social media platforms [188].
2. Algorithms to detect and flag any instances of inappropriate behavior or violations of professional ethical standards [50].

Social ethics

Refers to the moral principles and values that would involve considerations of how the use of social media affects the privacy, autonomy, and well-being of patients and other stakeholders, as well as issues related to fairness and equity [208]. For example, social ethics would require that healthcare providers and organizations respect the privacy of patients and protect their personal information when using social media platforms [58]. Social ethics would also require that healthcare providers and organizations take steps to ensure that the use of social media in healthcare does not create or reinforce existing health disparities, such as by providing access to care or health information only to certain groups of people who have access to social media. Overall, it provides a framework for evaluating the social and moral implications of using social media for healthcare interventions and for ensuring that these interventions are conducted ethically and responsibly. Several methods can contribute to the promotion of social ethics in AI:

1. Fairness-aware machine learning algorithms: These aim to mitigate unfairness in the training data and algorithmic decision-making process [154].
2. Privacy-preserving data analysis: These aim to protect sensitive data from unauthorized access, while still allowing for meaningful analysis [98].

3. Human-in-the-loop approaches: These incorporate human oversight and decision-making into the AI system, to ensure that the system is aligned with social values and ethical principles [54].

4. Explainable AI: This is a computational method that aims to make AI systems more transparent and understandable to users, so that they can make informed decisions about the ethical implications of the system's output [2].

5. Value-sensitive design: This seeks to identify and incorporate social values and ethical principles into the design and development of AI systems, in order to promote their alignment with social ethics [195].

Legal ethics

Refers to the ethical considerations related to complying with the laws, regulations, and policies surrounding healthcare data privacy and security [185]. This includes maintaining the confidentiality of patient data, adhering to informed consent and data-sharing agreements, and complying with relevant legal and ethical standards [92]. It also involves ensuring that the AI models used in social media intervention for healthcare are developed and used in compliance with relevant regulations and standards. Legal tools for ensuring ethics include:

1. HIPAA (Health Insurance Portability and Accountability Act): implementing privacy regulations for healthcare data [78].

2. GDPR (General Data Protection Regulation): complying with data protection laws and adhering to other relevant legal and regulatory frameworks that govern the use of AI in healthcare and social media interventions [174].

3. Ethical Research Board (ERB:)The idea of Ethics by Design suggests incorporating the services of an ERB while developing any product in an organization [108].

Table 2.6: Ethics evaluation metrics with mathematical formulation. *FP = False Positive, FN = False Negative, TP = True Positive, TN = True Negative*

Metric	Formula	Description
Bias [48]	$(\frac{\#FP\_for\_GroupA}{\#TN\_for\_GroupA})/(\frac{\#FP\_for\_GroupB}{\#TN\_for\_GroupB})$	Measures the extent to which an AI system exhibits bias towards a particular group or demographic.
Discrimination [156]	$\frac{P(positive\_outcome\_for\_GroupA)}{P(positive\_outcome\_for\_GroupB)}$	Measures whether an AI system is treating different groups of people unfairly.
Privacy [132]	$\frac{\#privacy\_violations}{\#individuals\_whose\_data\_was\_processed}$	Measures the extent to which an AI system is protecting the privacy of individuals.
Accountability [116]	$\frac{\#system\_was\_found\_to\_be\_at\_fault}{\#interactions\_with\_system}$	Measures whether an AI system can be held accountable for its actions.
Transparency [43]	$\frac{\#of\_decisions\_that\_can\_be\_explained}{Total\#of\_decisions\_made\_by\_AI\_system}$	Ensuring that the decision-making process of an AI system is clear and understandable to users.

2.6 FATE in Data Sets

Research has also been undertaken to improve transparency and accountability in the creation and use of datasets [84]. Here, the authors propose that thorough documentation should be kept for every step of the process, from the initial design to the final product. This would allow for clear identification of those who are responsible for any errors that may occur. Additionally, each stage of the development process should have a designated leader who takes ownership of their section of the program. Providing explanations for the purpose and function of each section can help to increase understanding and transparency. Furthermore, regular maintenance and documentation of updates should be conducted to not only minimize future mistakes but also boost morale among developers by highlighting progress made. The Adult Income dataset [51], the German Credit dataset [11], and the UCI Credit Card dataset [11] are commonly used while evaluating FATE models.

The performance of AI systems with respect to FATE principles can be evaluated using metrics to ensure that AI systems are making fair and unbiased decisions.

2.6.1 FATE toolkits

In recent years, researchers have been motivated to develop AI tools which can detect the level of bias present in a decision. Aequitas [170] produces reports that can facilitate equitable decision-making for policymakers and ML researchers, while the AI Fairness

360 [17] and Fairlearn [21] toolkit provides performance benchmarking for fairness algorithms. These libraries can therefore be used for assessing and mitigating bias in AI models, including methods for data pre-processing, model training, and post-processing.

It's important to note that FATE evaluations are not only a one-time assessment, but a continuous process, where metrics can be used to track the performance of the AI over time. Using these metrics, organizations can ensure that their AI systems are fair, accountable, transparent and ethical, and that they are making decisions that are in the best interest of all individuals.

2.7 Discussion

Medical corporations use social media to advertise their services, reach out to people and build a sense of community [71]. Social media platforms also provide a means for medical professionals to interact with patients and gather feedback, enabling them to improve their services. Social media can also be used to improve health through peer-to-peer encouragement, raise awareness on diseases, and for doctors to reach their patients through online consultations [33]. To combat misinformation, more fact-checking is needed, and more health institutions need to reach out to patients to ensure they are getting accurate information. The use of social media for health professionals should be carefully monitored to ensure patient confidentiality is maintained.

This study helps us identify the following principal findings:

- **RQ1**: What are the existing solutions to FATE (Fairness, Accountability, Transparency, and Ethics) when discussing healthcare on Social Media Platforms (SMPs)?

  The existing solutions to FATE when discussing healthcare on SMPs are:

  1. Healthcare fairness addressed through calibrated, statistical, and intersectional fairness. Calibrated fairness balances equal opportunities with personalized differences like language or location. Statistical fairness considers demographic information to avoid bias. Intersectional fairness considers multiple aspects of identity.

2. Accountability in healthcare on SMPs involves legal compliance, ethical principles in system design, technical functionality and privacy, and societal regulation and standardization. This involves protecting data privacy, avoiding discriminatory or unethical use of AI systems, conducting ethical impact assessments, promoting transparency, engaging stakeholders, conducting audits and evaluations, and holding decision-makers accountable.
3. Transparency in AI refers to understanding how an AI system works, including its algorithms, data sources, and decision-making processes. In social media for healthcare, transparency involves understanding how interventions are created, disseminated, and evaluated. Transparency is important for identifying and mitigating biases or errors, building stakeholder trust, and promoting engagement in social media-based health interventions.
4. Ethics in healthcare on SMPs involves developing fair, transparent, and beneficial AI technologies. This includes addressing privacy, data protection, bias, accountability, and explainability. Professional ethics and social ethics, such as patient privacy and autonomy, are also important. The goal is to promote ethical use of AI in healthcare on SMPs while minimizing risks and harms.

- **RQ2**: How do the different solutions identified in response to RQ1 compare to each other in terms of computational methods, approaches, and evaluation metrics?

  The different solutions identified in response to RQ1 for healthcare fairness on SMPs can be compared in terms of computational methods, approaches, and evaluation metrics. The solutions include: Calibrated, statistical, and intersectional fairness achieved through various computational methods such as data pre-processing, adversarial training, and decision trees/rules. Evaluation metrics include Equal Opportunity and Equal Odds. The solution for accountability involves regulatory measures and public-private partnerships to ensure transparency, fairness, privacy, and accountability. Transparency in AI can be achieved through algorithmic transparency, data transparency, and process transparency. Algorithmic transparency can be increased through methods like feature importance analysis, model interpretability, and explanation generation. Data transparency can be improved through data visualization, profiling, and lineage analysis. Process transparency can be enhanced through data provenance tracking, interpretability, explanation generation, auditability, monitoring, and open-source development. Ethics in healthcare on SMPs can be promoted through simulation, modeling, data analytics, sentiment

analysis, crowdsourcing, and automated systems, while also considering professional ethics and social ethics.

- **RQ3**: What is the strength of evidence supporting the different solutions?

  The strength of evidence supporting different solutions for healthcare fairness on Social Media Platforms (SMPs) varies depending on factors such as research quality, methodology, and statistical significance. Calibrated fairness, statistical fairness, and intersectional fairness have established concepts with significant research support. Computational methods like data pre-processing, adversarial training, and decision trees/rules are commonly used, but evidence of their effectiveness may vary. Evaluation metrics such as Equal Opportunity and Equal Odds are commonly used but rely on established statistical measures. Ethics in healthcare on SMPs, including privacy protection and bias mitigation, are guided by established principles, but evidence supporting specific solutions may vary. Solutions like simulation, modeling, data analytics, and crowdsourcing are widely used, but their evidence may vary depending on context. Consulting reputable sources for up-to-date research findings is important due to the dynamic nature of the field.

2.8 Limitations and Future Research Directions

This study offers a comprehensive overview of the challenges and progress related to FATE in AI. Despite advancements, challenges remain for AI systems in healthcare, including ethical considerations for patient decision-making, accuracy, and understanding of decision-making processes. While this study focused on searching the Google Scholar database, I did not consider other resources such as Web of science, IEEExplore, ACM Digital Library, and grey literature. Excluding them may have resulted in some important studies not being a part of this study. Obtaining trustworthy data sets and informed user consent, especially for large language models like ChatGPT, which have the potential of being used in clinical settings, is challenging. Overconfidence in AI systems can also lead to skepticism from clinicians. Additionally, the lack of mathematical formulation for many FATE computational methods and approaches creates a gap between computational and evaluation metrics.

2.9 Conclusion

The purpose of this review was to provide a comprehensive analysis of FATE solutions in AI for social media and healthcare, and to highlight recent trends and research gaps in the field. By examining the definitions, computational methods, approaches, and data sets used in the literature, we identified both the progress made and the challenges that remain in achieving FATE in AI. Through our evaluation of the papers, we also highlighted the need for researchers to use appropriate evaluation metrics and data sources when analyzing their approaches. While some progress has been made, there is still much work to be done in order to address the remaining challenges. We hope that this review will serve as a useful resource for researchers and stakeholders, and that it will encourage further research in this important area. Ultimately, our goal is to support the development of FATE-ready AI systems that can be deployed ethically and responsibly in social media and healthcare.

Chapter 3

Synergy Between Public and Private Health Care Organizations During COVID-19 on Twitter: Sentiment and Engagement Analysis Using Forecasting Models

All of this chapter was published in the JMIR Medical Informatics (2022) as the following peer-reviewed article [177]:

• Singhal A, Baxi MK, & Mago V. Synergy Between Public and Private Health Care Organizations During COVID-19 on Twitter: Sentiment and Engagement Analysis Using Forecasting Models

Using the advancements in the field of Natural Language Processing, this chapter proposes an approach for organizations to structure their future Twitter content to ensure maximum user engagement.

Keywords : social media, health care, Twitter, content analysis, user engagement, sentiment forecasting, natural language processing, public health, pharmaceutical, public engagement

3.1 Introduction

3.1.1 Background

Social media platforms (SMPs), such as Twitter, Facebook, and Reddit, are commonly used by people to access health information. In the United States, 8 in 10 internet users access health information online, and 74% of these use SMPs. Meanwhile, public health agencies and pharmaceutical companies often use social media to engage with the public [198]. SMPs significantly contribute to the community by providing a communication platform for the public, patients, and health care professionals (HCPs) to talk about health concerns, eventually leading to better outcomes [38]. Additionally, SMPs also function as a medium to motivate patients by promoting health care education and providing the latest information to the community [198]. Analyzing social media content in the health care domain can reveal important dimensions, such as audience reach (eg, followers and subscribers), post source (eg, pharmaceutical companies, public health agencies), and post interactivity (eg, number of likes, retweets) [221]. A recent study discussed a machine learning (ML) approach to examining COVID-19 on Twitter [211]. Although it identifies discussion themes, there is no research on understanding the content shared by public health agencies and private organizations.

3.1.2 Related Works

The positive impacts of using SMPs by patients and HCPs have been previously discussed [18]. Patients feel empowered and develop positive relationships with their HCPs. For instance, Ventola [198] discussed SMPs as a tool to share and promote healthy habits, share information, and interact with the public. Li et al [112] presented an analysis of social media's impact on the public. Their research discusses public perceptions of health-related content being classified as true, debatable, or false; the study shows that people have a strong tendency to adopt collective opinions while sharing health-related statements on social media.

There are different topic-clustering and content analysis techniques available to identify the characteristics of stakeholders (eg, pharmaceutical companies' tweets for drug

information) on SMPs [61,119,138,194]. A previous study presented an overview of techniques used for sentiment analysis in health care [1]. The researchers discuss multiple lexicon-based and ML-based approaches. The previous discussion on pharmaceutical companies has focused on COVID-19 vaccine–related public opinions [32, 159]. Using latent dirichlet allocation (LDA) and valence aware dictionary and sentiment reasoner (VADER), researchers have examined topics, trends, and sentiments over time [32].

Prior research work has also focused on the response of G7 leaders during COVID-19 on Twitter [77, 167]. The research classified viral tweets into appropriate categories, the most common being informative. Furthermore, researchers have recently presented a discussion on the harms and benefits of using Twitter during COVID-19 [165]. An epidemiological study conducted in 2020 investigated the news-sharing behavior on Twitter. Although it concluded that tweets that include news articles sharing pandemic information are popular, they cannot substitute public health agencies, organizations, or HCPs [151]. In addition, the study of public sentiments via artificial intelligence (AI) can provide a way to frame public health policies [83].

COVID-19 led to a rapid change in public sentiments over a short span of time [121]. People expressed sentiments of joy and gratitude toward good health and sadness and anger at the loss of life and stay-at-home orders [53, 121]. Understanding public perceptions toward health-related content is important. Although the majority of people have a positive attitude toward social media, some feel more attention is required to promote the credibility of shared information [31,59,60,173]. Attempts have been made to capture peoples' reactions to the pandemic; however, they are limited in scope. One study investigated the concerns originating toward public health interventions in North America via topic modeling [88], while another examined the role of beliefs and susceptibility information in public engagement on Twitter [190]. Statistical analysis also shows that health care organizations have to come forward to engage more with consumers [104]. The importance of risk communication strategies while using SMPs cannot be undermined [180].

Although a tweet's engagement and sentiment can only be calculated once it has been posted, forecasting presents a fascinating way to predict the sentiments beforehand. Time series–based strategies, such as autoregressive integrated moving average (ARIMA) and vector autoregressions (VAR), have been used for forecasting emotions from SMPs [128,193]. The seasonal autoregressive integrated moving average with exoge-

nous factors (SARIMAX) model was recently used to gain insights into people's current emotional state via sentiment nowcasting on Twitter [135].

ML and natural language processing (NLP) algorithms have been recently used in various instances; for example, Bayesian ridge and ridge regression models were used for emotion prediction and health care analysis on large-scale data sets [45,79]. The elastic net and lasso regression have been previously used for health care access management and information exchange [56, 114], while linear regression, decision tree, and random forest models are commonly used for epidemic-level disease tracking [176]. Different regression boosting algorithms, such as AdaBoost, light gradient boost, and gradient boost, have also been used for disease outbreak prediction [176]. Prophet, a Python library package, was recently used for COVID-19 outbreak prediction [134].

3.1.3 Objective

The implications of social media communication by HCPs have been extensively discussed [46,142]. Although they focus on the advantages and methods of extracting health- and disease-related content from social media, there is currently a lack of understanding of how social media usage by public health agencies, nongovernment organizations (NGOs), and pharmaceutical companies resonates with society. Additionally, the study of tweets' sentiments can supplement existing models for generating content for future tweets. Predicting the tweet sentiment is 1 way to achieve this goal. Therefore, it is crucial to convert this textual content into information for formulating future strategies and gaining valuable insights into perceptions of social media users.

The remainder of the paper is structured as follows: First, a preliminary analysis of topic modeling using the best-performing clustering algorithm is presented in the Methods section, followed by sentiment and engagement analysis using CardiffNLP's twitter-roberta-base-sentiment model. We then conducted time series–based sentiment forecasting using 16 univariate models on the complete data set. The Results section outlines model topics obtained, which were used for generating heatmaps to obtain insights into topicwise tweets. Next, we discussed user engagement with its impact to understand whether there were specific occurrences of higher levels of engagement impacted by any offline events. In addition, we discussed results from best-performing sentiment-

forecasting models. Finally, in the Discussion section, we draw conclusions and present an outline for future work.

3.2 Methods

3.2.1 Data Set

The data for this study (181,469 tweets) were gathered from the accounts of major US and Canadian health care organizations, pharmaceutical companies, and the World Health Organization (WHO) using the Twitter Academic API for Research v2 [1] during the time frame of January 1, 2017, to December 31, 2021. The top 5 pharmaceutical companies were selected based on the recommendations made by HCPs on Twitter [2]. Table 3.1 lists the number of tweets scraped for each Twitter handle. Each organization is referred to as a user, and the type of organization (ie, pharmaceutical company, public health agency, NGO) is referred to as a user group for the scope of this study.

The complete timeline was divided into 2 phases for analysis, before COVID-19 and during COVID-19, based on the confirmation of the first COVID-19 community transmission case in North America on February 26, 2020 [39]. Figure 3.1 presents an overview of the research framework.

3.2.2 Content Analysis

The content of each user was divided into 2 phases, before and during COVID-19. We performed topic modeling on the tweets authored by the organizations by using the topics yielded by the best-performing topic model in order to explore the most and least talked about topics with the help of heatmaps. Additionally, we examined the top 10 hashtags used by these organizations.

[1] https://developer.twitter.com/en/products/twitter-api/academic-research

[2] https://creation.co/knowledge/hcps-discuss-booster-shot-to-decrease-the-high-spread-of-the-delta-variant/

Table 3.1: Distribution of tweets for the selected user accounts of 3 types of organizations.

Name of organization (Twitter handle)	Before COVID-19, n (%)	During COVID-19, n (%)	Total tweets, N
Public health agencies			
Centers for Disease Control and Prevention (CDCgov)	8435 (58.6)	5963 (41.4)	14,398
Centers for Disease Control and Prevention (CDC_eHealth)	1376 (86.3)	219 (13.7)	1594
Government of Canada for Indigenous (GCIndigenous)	3505 (54.0)	2989 (46.0)	6494
Health Canada and PHAC (GovCanHealth)	7878 (17.2)	37,907 (82.8)	45,785
US Department of Health & Human Services (HHSGov)	7890 (56.9)	5969 (43.1)	13,859
Indian Health Service (IHSgov)	1090 (44.7)	1346 (55.3)	2436
Canadian Food Inspection Agency (InspectionCan)	4145 (62.2)	2516 (37.8)	6661
National Institutes of Health (NIH)	5837 (71.6)	2314 (28.4)	8151
National Indian Health Board (NIHB1)	1247 (51.1)	1195 (48.9)	2442
US Food and Drug Administration (US_FDA)	5810 (59.7)	3925 (40.3)	9735
Total	47,213 (42.3)	64,343 (57.7)	111,555
Pharmaceutical companies			
AstraZeneca (AstraZeneca)	3462 (78.2)	963 (21.8)	4425
Biogen (biogen)	1819 (61.9)	1120 (38.1)	2939
Glaxo SmithKline (GSK)	4200 (69.3)	1857 (30.7)	6057
Johnson & Johnson (JNJNews)	4813 (71.4)	1926 (28.6)	6739
Pfizer (pfizer)	3637 (64.1)	2039 (35.9)	5676
Total	17,931 (69.4)	7905 (30.6)	25,836
NGO (Non Government Organization			
World Health Organization (WHO)	24,775 (56.2)	19,303 (43.8)	44,078

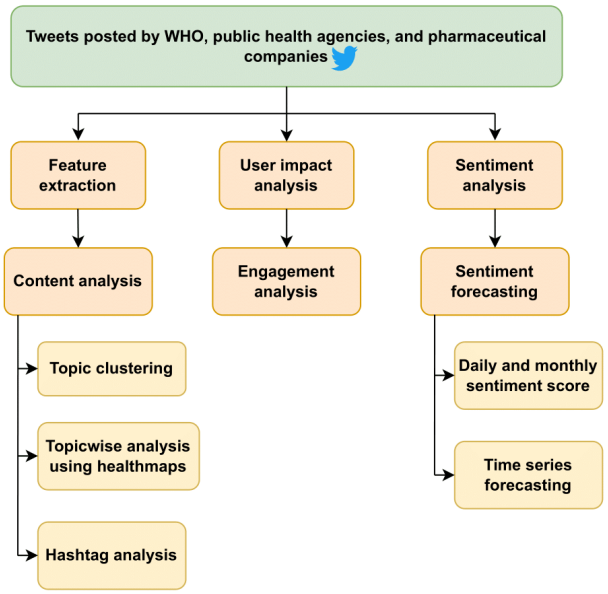


Figure 3.1: Overall research framework. WHO: World Health Organization.

3.2.3 Preprocessing

First, all nonalphabets (numbers, punctuation, new-line characters, and extra spaces) and Uniform Resource Locators (URLs) were removed using the regular expression module (re 2.2.1)[3] for all tweets. The cleaned text was then tokenized using the nltk 3.2.5 library[4]. Next, stopwords were removed, followed by stemming using PorterStemmer, and lemmatizing using the WordNetLemmatizer from nltk.

3.2.4 Topic Modeling

Researchers have used term frequency–inverse document frequency (TF-IDF) to create document embeddings for tweets [115]. Following their approach, we preprocessed and generated document embeddings for tweets and input them to 5 different clustering algorithms: LDA, parallel LDA, nonnegative matrix factorization (NMF), latent semantic indexing (LSI), and the hierarchical dirichlet process (HDP). These clustering algorithms were executed 5 times with varying random seed values. The seed values accounted for the short and noisy nature of tweets. We calculated the coherence scores of the topic models, c_{umass} [93] and c_v [164], to confirm performance consistency over multiple runs.

We used Gensim LDA[5], Gensim LDA multicore (parallel LDA)[6], and Gensim LSI [7] models. For NMF and HDP models, we used online NMF[8] for large corpora and online variational inference[9] models, respectively.

3.2.5 Heatmaps

Heatmaps were generated using seaborn to analyze the volume of tweets for each topic. The topics yielded by the best-performing topic model as per the time phase (ie, before and during COVID-19) were leveraged to generate heatmaps. Each cell represented the

[3]https://pypi.org/project/regex/
[4]https://pypi.org/project/nltk/
[5]https://radimrehurek.com/gensim/models/ldamodel.html
[6]https://radimrehurek.com/gensim/models/ldamulticore.html
[7]https://radimrehurek.com/gensim/models/lsimodel.html
[8]https://radimrehurek.com/gensim/models/nmf.html
[9]https://radimrehurek.com/gensim/models/hdpmodel.html

total count of tweets for a particular topic by an organization. For example, among pharmaceutical companies, AstraZeneca had the highest number of tweets (n=1729, 49.9%) before COVID-19 for chronic diseases.

3.2.6 Hashtags

The top 10 hashtags mentioned in the users' tweets were evaluated using the advertools 0.13.0 module[10]. This tool extracts hashtags in social media posts. It was used for analyzing the similarities and differences in the tweeting behavior before and during COVID-19 and conducting topic analysis.

3.2.7 Sentiment Analysis

Sentiment analysis is an NLP approach used to categorize the sentiments appearing in Twitter messages based on the keywords used in each tweet. We tested different models that classify a user's tweet in 1 of 3 categories: positive, negative, and neutral. Although there is no common threshold for how many tweets should be sampled, we witnessed a range of around 2000 tweets [7, 155, 155] to several thousand tweets [69, 140, 168] when testing a model. For this study, we sampled 3000 tweets uniformly distributed over the span of our data collection time frame and from all Twitter handles. The tweets were then labeled by 3 distinct annotators, and the sentiment category with the highest votes was chosen as the overall sentiment. CardiffNLP's twitter-roberta-base-sentiment model[11], which is trained on a 60 million Twitter corpus, was used to obtain sentiment labels on the sampled data set. We checked for similarity between human annotations and model labels, and the similarity percentage for CardiffNLP's model was 69.96%; the model was therefore used to predict the sentiment on the remaining tweets of the users.

[10]https://pypi.org/project/advertools/
[11]https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

3.2.8 Engagement Analysis

For a given user, Twitter defines the engagement rate[12] as presented in Equation 3.1:

$$EngagementRate = \frac{Engagement}{Impressions} * 100 \tag{3.1}$$

where "Engagement is the summation of the number of likes, replies, retweets, media views, tweet expansion, profile, hashtag, URL clicks, and new followers gained for every tweet, and Impressions is the total number of times a tweet has been seen on Twitter, such as through a follower's timeline, Twitter search, or as a result of someone liking your tweet."

Researchers have analyzed the impact (popularity) of Twitter handles by proposing heuristic and neural network–based models [42, 162, 183]. We defined it as a function of followers, following, the total number of tweets, and the profile age and calculated it using Equation 3.2:

$$Impact_{\text{user}} = \frac{(followers * listedCount) * \log_{10}(\frac{followers}{following} + 1)}{tweetCount * profileAge} \tag{3.2}$$

where listedCount is the number of public lists of which this user is a member.

The total number of tweets produced by a user was considered inversely proportional to the user's impact, because a user tweeting occasionally and receiving higher engagement is more impactful than a user tweeting regularly with lower engagement.

Engagement analysis was performed to quantify the popularity of a topic generated. The engagement for each user was defined as the product of average engagement per day and their impact, as described in Equation 3.3. The average engagement per day was calculated as the sum of the count of likes, replies, retweets, and quotes per day. These reactions were aggregated from January 1, 2017, to December 31, 2021.

[12]https://help.twitter.com/en/managing-your-account/using-the-tweet-activity-dashboard

$$AvgEngagement/day = \frac{likes + replies + retweets + quotes}{4 * tweetsPerDay} * Impact_{\text{user}} \quad (3.3)$$

The exponential moving average (EMA) was calculated with a window span of 151 days for every user, and outliers were removed using the z-score, followed by smoothening of the average engagement per day to the eighth degree using the Savitzky-Golay filter [125].

3.2.9 Sentiment Forecasting

To forecast the sentiment per day, we first needed to quantify the overall sentiment of the tweets from each user every day. We leveraged CardiffNLP's twitter-roberta-base-sentiment model[13] to calculate the sentiments of all the tweets collected for our analysis and then calculated the daily sentiment score, as mentioned in Equation 3.4, based on the sentiment category with the maximum number of tweets for that day, followed by assigning the sentiment score based on the sentiment: 0 for neutral sentiment, the ratio of the count of positive tweets to total tweets for positive sentiment, and the negation of the ratio of the count of negative tweets to the total tweets for negative sentiment.

$$dailySentimentScore = \begin{cases} 0 & : maxSentiment(tweets) = neutral \\ \frac{count(PositiveTweets)}{totalTweets} & : maxSentiment(tweets) = +ve \\ -\frac{count(NegativeTweets)}{totalTweets} & : maxSentiment(tweets) = -ve \end{cases} \quad (3.4)$$

The daily sentiment scores were then resampled to a monthly mean sentiment score, which also helped us in handling missing values, if any. The complete timeline was divided into 2 phases (ie, before and during COVID-19), as discussed before, and the sentiment score was forecasted on 20% of the data set in each period for all user groups.

A grid search was used to find optimal hyperparameters, and 5-fold cross-validation was performed for every model. The statsmodel library[14] was used for ARIMA[15] and

[13]https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment
[14]https://www.statsmodels.org/stable/index.html
[15]https://www.statsmodels.org/devel/generated/statsmodels.tsa.arima.model.ARIMA.html

SARIMAX[16] models, and pycaret[17] was used for regression-based models. We also reported the performance of the prophet[18] model on the data set.

Three metrics, the mean absolute error (MAE), the mean square error (MSE), and the root-mean-square error (RMSE), were selected to evaluate the forecasting accuracy of the models. We considered 1-step-ahead forecasting for this study as it helped avoid problems related to cumulative errors from the preceding period.

3.2.10 Computational Resources

The study was performed using Compute Canada (now called the Digital Research Alliance of Canada) resources, which provide access to advanced research computing (ARC), research data management (RDM), and research software (RS). The following is a list of the computing resources offered by one of the clusters from National Services (Digital Research Alliance), Graham:

- Central processing unit (CPU): 2x Intel E5-2683 v4 Broadwell@2.1 GHz
- Memory (RAM): 30 GB

3.3 Results

3.3.1 Content Analysis

The details of the parameters used for each model are discussed in Appendix Table 6.1. Table 3.2 shows the mean coherence scores (cv and cumass) for each clustering algorithm. Although the HDP had the highest cv scores in both time phases (ie, 0.696 and 0.650 before and during COVID-19, respectively), NMF had the best cumass scores (–3.653 and –3.794, respectively) and generated the most meaningful topics for the data set (see Appendix Tables 6.2 and 6.3). Therefore, the top 5 topics generated by NMF were selected

[16]https://www.statsmodels.org/devel/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html
[17]https://pypi.org/project/pycaret/
[18]https://pypi.org/project/prophet/

Table 3.2: Mean coherence scores and CPU time for different clustering algorithms.

Clustering algorithm	cv	cumass	Time taken (minutes:seconds)
Before COVID-19			
LDA	0.352	–5.526	17:11
Parallel LDA	0.396	–3.709	5:48
NMF	0.493	–3.653	7:38
LSI	0.316	–5.921	0:16
HDP	0.696	–18.668	3:24
During COVID-19			
LDA	0.456	–5.688	14:01
Parallel LDA	0.446	–3.990	6:08
NMF	0.567	–3.794	7:04
LSI	0.381	–5.356	0:16
HDP	0.650	–17.610	3:01

to search for on the first page of Google Search results. The resulting contents were then retrieved to interpret the extracted topic keywords to propose a suitable topic name. For example, for the set of keywords yielded by the topic model "community health, care, community health services, health center, family health centers, community plan, community clinic, family health care, qualified health centers, health services," we assigned the topic community health care.

The scaled heatmaps showing the topic distribution for different Twitter handles are shown in Figure 3.2. Prior to COVID-19, chronic diseases were the most active topic, with a total of 9488 tweets from pharmaceutical companies and WHO (see Figure 3.2a). However, during COVID-19, we observed that COVID-19, health research, and chronic diseases were the most-discussed topics, with 52,148 tweets from all data sets combined (see Appendix Figure 6.1).

This shift in the tweets' content was observed across the complete data set, and we further made the following inferences:

- Before COVID-19: Chronic diseases were the most talked about topic for pharmaceutical companies (AstraZeneca, 1729, 49.9%, tweets; Pfizer, 1168, 32.1%, tweets) and for WHO (4831, 19.5%, tweets), followed by tweets on health research (WHO, 1703, 6.9%, tweets; AstraZeneca, 1037, 29.9%, tweets). This is supported by Figure 3.3a, which shows cancer, lungcancer, alzheimers, hiv, and ms to be prominently

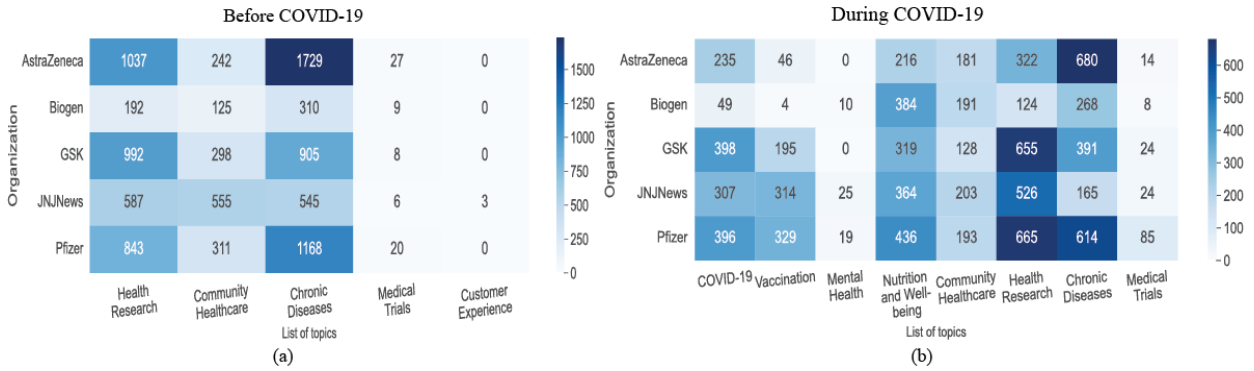


Figure 3.2: Scaled heatmaps showing topic distribution for pharmaceutical companies before and during COVID-19.

used in tweets. Among public health agencies, the NIH's and the CDC's Twitter handles were the most active, with 1840 (31.6%) and 1742 (20.6%) tweets discussing health research and chronic diseases, respectively, strongly supported by the most used hashtags nativehealth and foodsafety (refer to Appendix Figure 6.2).

- During COVID-19: Chronic diseases and health research were the most active topics for AstraZeneca (680, 70.6%, tweets) and Glaxo SmithKline (GSK, 655, 35.2%, tweets), respectively. In addition, COVID-19 and vaccination were most talked about by GSK (398, 21.4%, tweets) and Pfizer (396, 19.4%, tweets). Figure 3.3b shows the hashtags supporting this: covid19, alzheimers, cancer, multiplesclerosis, and vaccine. GovCanHealth was by far the most active public health agency on Twitter, with 16,832 (87.2%) tweets on health research, 16,449 (85.2%) tweets on vaccination, and 14,260 (73.8%) tweets on COVID-19, having covid19, coronavirus, and covidvaccine as trending hashtags. The majority of the tweets by WHO were on COVID-19 (8911 tweets) and vaccination (2131 tweets), with covid19, coronavirus, and vaccineequity appearing frequently in the tweets (refer to Appendix Figure 6.2).

3.3.2 Engagement Analysis

WHO (user impact=4171.24) had the highest impact overall, followed by public health agencies (CDC user impact=2895.87; NIH user impact=891.06). Among pharmaceutical companies, Pfizer's user impact was the highest at 97.79. The user impact was normalized between the range of 0 and 1 and is shown in Figure 3.4.

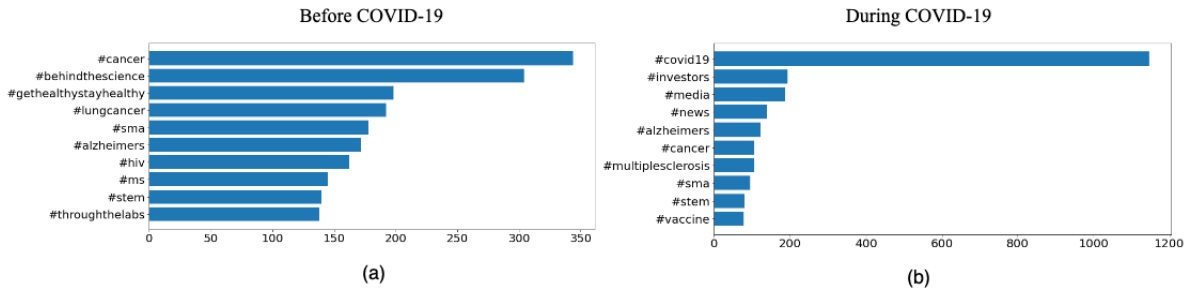


Figure 3.3: Top hashtags of pharmaceutical companies before and during COVID-19.

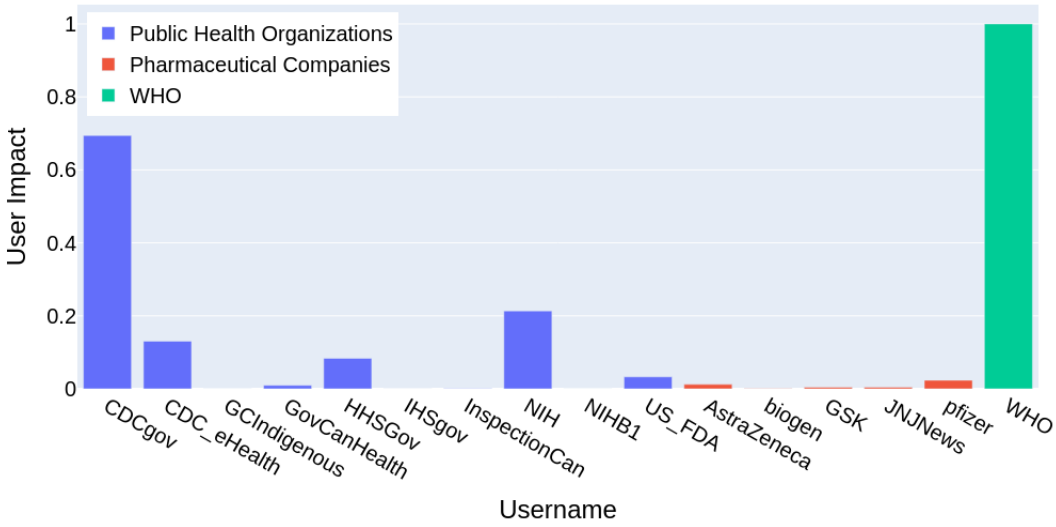


Figure 3.4: User impact of all Twitter handles scaled between 0 and 1. CDC: Centers for Disease Control and Prevention; NIH: National Institutes of Health; WHO: World Health Organization.

3.3.3 Engagement Analysis

Among pharmaceutical companies, Pfizer's user engagement was far higher than that of others (Figure 3.5), both before and during COVID-19, with the highest engagement observed at the time of its COVID-19 vaccine's success in November 2020. A jump in engagement was also observed in May 2021, when Pfizer announced its plan for helping India fight the second wave of coronavirus (refer to Appendix Table 6.4).

A similar trend was observed in public health agencies, with the CDC's account showing the highest user engagement between March and June 2020, the early months of the COVID-19 pandemic. A sharp rise in user engagement was observed in May 2021, when

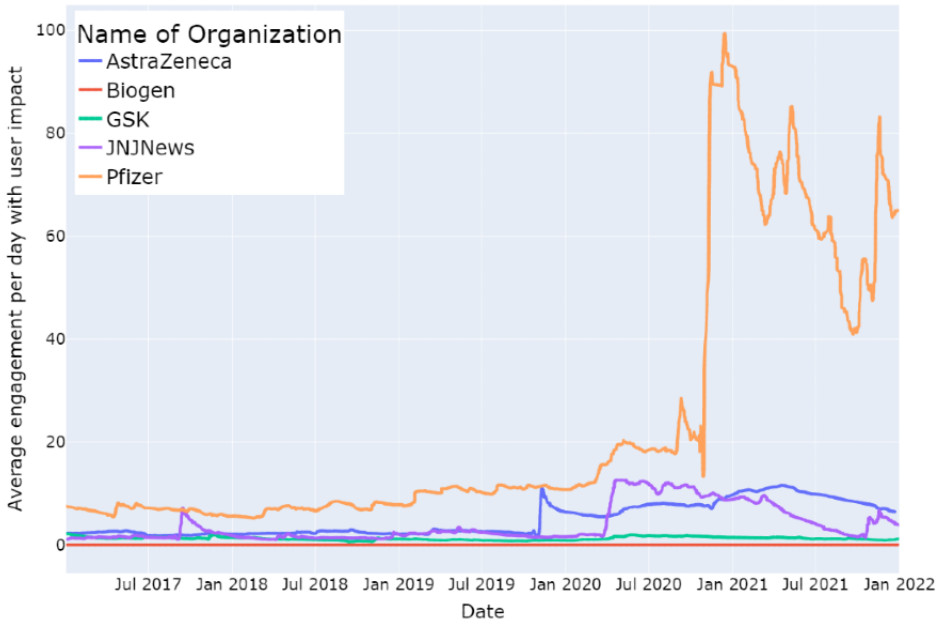


Figure 3.5: User engagement on Twitter accounts of pharmaceutical companies from January 1, 2017, to December 31, 2021.

the CDC announced a relaxation on social distancing and masking rules for fully vaccinated individuals. The user engagement on WHO's account varied significantly over time. Its engagement was the highest in the time frame of February-April 2020, the early months of the pandemic, similar to what was observed for public health agencies. A sharp increase was seen in October 2020 following the announcement of the World Mental Health Day and in late 2020, when WHO made an announcement for COVID-19 vaccine development (refer to Appendix Figure 6.3).

3.3.4 Sentiment Forecasting

Table 3.3 shows the MAE, MSE, and RMSE for the 16 models used on the data sets. Overall, ARIMA (univariate) and SARIMAX models performed best on the majority of the subsets of the data (divided as per the organization and period), and we further made the following inferences:

- Before COVID-19: ARIMA and SARIMAX models generated the lowest MSE (0.005) and RMSE (0.072) for pharmaceutical companies. When measuring the model per-

Table 3.3: Results of time series sentiment forecasting using different ML models (all metrics are 5-fold cross-validation).

	Pharmaceutical companies						Public health agencies						WHO					
	Before COVID-19			During COVID-19			Before COVID-19			During COVID-19			Before COVID-19			During COVID-19		
Models	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
ARIMA	0.063	0.005	0.072	0.098	0.013	0.112	0.027	0.001	0.032	0.240	0.082	0.286	0.066	0.006	0.080	0.106	0.012	0.111
SARIMAX	0.065	0.005	0.072	0.084	0.011	0.104	0.028	0.001	0.031	0.709	0.011	0.106	0.054	0.004	0.061	0.047	0.004	0.066
Bayesian ridge	0.083	0.010	0.100	0.102	0.018	0.119	0.031	0.001	0.037	0.141	0.037	0.163	0.075	0.009	0.087	0.061	0.008	0.075
Ridge regression	0.069	0.008	0.085	0.079	0.011	0.094	0.030	0.002	0.038	0.124	0.029	0.147	0.076	0.009	0.091	0.056	0.007	0.068
CatBoost regressor	0.066	0.007	0.080	0.072	0.008	0.086	0.027	0.001	0.035	0.104	0.023	0.127	0.079	0.009	0.089	0.052	0.007	0.065
K-neighbors regressor	0.070	0.009	0.087	0.075	0.008	0.087	0.030	0.001	0.036	0.093	0.022	0.113	0.081	0.011	0.100	0.050	0.007	0.061
Elastic net	0.070	0.008	0.088	0.080	0.009	0.093	0.029	0.001	0.035	0.087	0.021	0.109	0.082	0.011	0.100	0.046	0.006	0.059
Lasso regression	0.070	0.008	0.088	0.080	0.009	0.093	0.029	0.001	0.035	0.087	0.021	0.109	0.082	0.011	0.100	0.046	0.006	0.059
Random forest regressor	0.065	0.007	0.081	0.080	0.010	0.093	0.028	0.001	0.034	0.110	0.024	0.134	0.082	0.009	0.090	0.047	0.006	0.060
Light gradient boosting machine	0.070	0.008	0.088	0.080	0.009	0.093	0.029	0.001	0.035	0.087	0.021	0.109	0.082	0.011	0.100	0.046	0.006	0.059
Gradient boosting regressor	0.075	0.008	0.086	0.079	0.010	0.094	0.029	0.001	0.036	0.141	0.034	0.168	0.082	0.010	0.094	0.051	0.008	0.064
AdaBoost regressor	0.070	0.007	0.082	0.080	0.010	0.091	0.029	0.001	0.037	0.084	0.020	0.105	0.087	0.010	0.096	0.057	0.007	0.072
Extreme gradient boosting	0.068	0.009	0.087	0.080	0.011	0.098	0.031	0.002	0.040	0.151	0.045	0.171	0.087	0.011	0.098	0.055	0.007	0.065
Decision tree regressor	0.076	0.009	0.086	0.087	0.013	0.106	0.029	0.001	0.037	0.112	0.030	0.142	0.098	0.014	0.111	0.048	0.006	0.061
Linear regression	0.245	0.312	0.314	0.094	0.017	0.114	0.157	0.164	0.216	0.124	0.029	0.148	2.367	52.719	3.334	0.062	0.008	0.076
Prophet	0.108	0.016	0.126	0.089	0.011	0.104	0.040	0.002	0.049	0.120	0.015	0.124	0.114	0.020	0.143	0.086	0.011	0.106

formance through the MAE, ARIMA performed better than all other models (0.063). A similar trend was observed for public health agencies, with ARIMA having the lowest MAE (0.027) and SARIMAX having the lowest RMSE (0.031) and a tie between them for the MSE (0.001). SARIMAX had the lowest MAE (0.054), MSE (0.004), and RMSE (0.080) on the WHO data set.

- During COVID-19: Using the CatBoost regressor gave the lowest MAE (0.072) and RMSE (0.086), while the K-neighbors regressor yielded the lowest MSE (0.008) for pharmaceutical companies. Performing regression using AdaBoost generated the lowest MAE (0.084) and RMSE (0.105) among all models used, and SARIMAX had the lowest MSE (0.011) for public health agencies. For WHO, the elastic net, lasso regression, and light gradient boosting performed equally well, with all 3 models having the same MAE (0.046) and RMSE (0.059), and SARIMAX had the lowest MSE (0.004).

Figure 3.6a shows the 1-step-ahead forecast for pharmaceutical companies before COVID-19 using ARIMA. The model was trained on sentiment scores from January 2017 to June 2019 and tested on data from July 2019 to February 2020 for tweets before COVID-19. The 1-step-ahead forecasting aligned well with the observed sentiment scores, and we obtained similar results for public health agencies and WHO. The organizations showed some deviations from observed sentiments while conducting 1-step-ahead forecasting during COVID-19, making it difficult to predict their sentiment accurately, as seen in Appendix Figure 6.4.

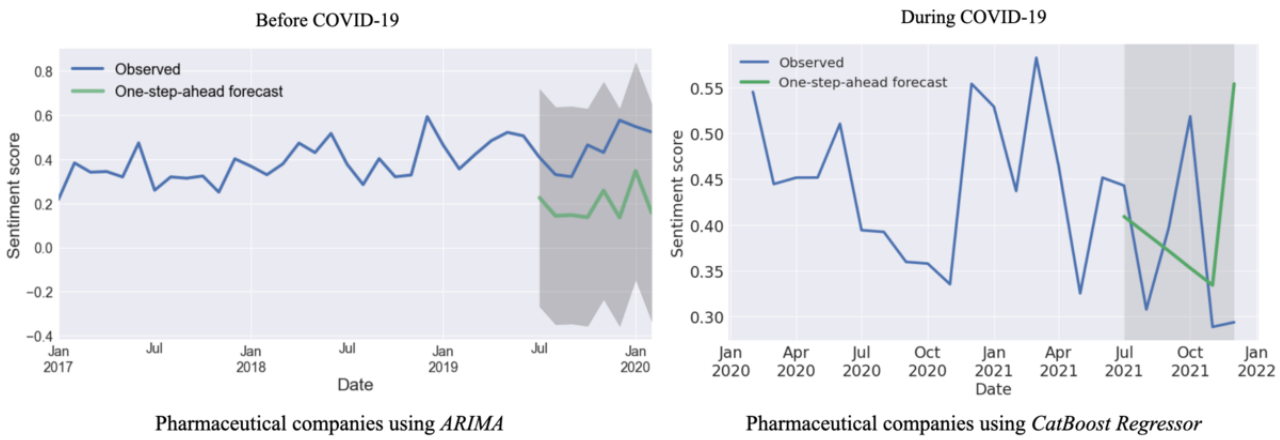


Figure 3.6: One-step-ahead forecast for all pharmaceutical companies before and during COVID-19 using the best-performing models from Appendix Table 6.1). ARIMA: autoregressive integrated moving average.

To verify the forecasting performance of these models, we checked for the nature of their residual errors (ie, whether the residuals of the models were normally distributed with mean 0 and SD 1 and were uncorrelated). From Appendix Figure 6.5, as in the case of public health agencies, before COVID-19 using ARIMA, we confirmed the aforementioned through plot_diagnostics. The green kernel density estimation (KDE) line closely followed the normal distribution (N $\in$ 0,1) line in the top-right corner of Appendix Figure 6.5, which is a positive indicator that the residuals were scattered normally. The quantile-quantile (Q-Q) plot on the bottom left shows that the distribution of residuals (blue dots) approximately followed the linear trend of samples drawn from a standard normal distribution, N. This confirms again that the residuals were normally distributed. The residuals over time (top left in Appendix Figure 6.5) showed no apparent seasonality and have 0 mean. The autocorrelation plot (ie, correlogram) attested this, indicating that the time series residuals exhibited minimal correlation with lagged forms of themselves. Thus, these findings encouraged us to believe that our models provide an adequate fit, which might aid us in understanding the sentiments of the organizations and forecasting their values without overburdening our hardware with computationally heavy models.

3.4 Discussion

3.4.1 Principal Findings

In this paper, we proposed a framework for using NLP-based text-mining techniques for performing comprehensive social media content analysis of various health care organizations. We processed reasonably large amounts of textual data for topic modeling, sentiment and engagement analysis, and sentiment forecasting. Our study revealed the following key findings:

- Being the most active organization on social media does not translate to more user impact. WHO and the US public health agency CDC generated far more user impact than the Public Health Agency of Canada, even though the latter had a high number of relevant tweets when analyzed topicwise. People are more likely to engage with neutral tweets, which usually consist of some public health announcement rather than exclusively positive or negative tweets. This might mean that organizations can leverage this knowledge while creating content for social media posts in the future to increase their visibility in the online sphere.
- Certain topics normally translate to more user engagement. Although the content on chronic diseases and health research dominated most of the tweets posted over the study period, there was a marked shift toward a discussion on COVID-19 and vaccination for public health agencies, more than what was observed in pharmaceutical companies. Tweets on COVID-19 and chronic diseases generate more interest among the public. Perhaps surprisingly, we found that people are not much receptive to content on medical trials, often shared by pharmaceutical companies, unless it concerns a public health emergency, such as the COVID-19 pandemic. Using particular hashtags certainly helps in generating engagement, as we found that most user engagement was highly skewed toward tweets concerning COVID-19. Moreover, our study revealed that compared to the user engagement patterns found in the majority of health care organizations (ie, with peaks observed around major events or announcements), there are wide variations in user engagement for WHO. This could be due to the global presence of WHO, implying that it might not be the same set of followers engaging with its content every time, but rather only those who are impacted by or interested in the content in some way.

- When the content is structured, results tend to exceed expectations. We conducted sentiment forecasting on the data sets using different moving averages and various ML univariate models. Surprisingly, we observed that when the content is structured, as is normally the case for that available on official Twitter accounts, results tend to exceed expectations, more so before COVID-19 than during COVID-19. The models used in this research are able to predict monthwise tweet sentiment with high accuracy and low errors. This helped us in analyzing our work in-depth, and we did not need to create any multivariate ML models. Results show that commonly used ARIMA and SARIMAX models work well, and they can be used for predicting tweet sentiments on live data. This could also help organizations correlate tweet sentiment with user engagement. For example, the highest engagement on Pfizer's tweets was for the ones labeled neutral, implying that the organization should structure the content of its future tweets in a similar manner to maintain higher levels of engagement. Furthermore, tweets that mention more news-relevant content might be able to translate it into more user engagement.

3.4.2 Limitations and Future Work

There are some limitations of this study that could be addressed in future research. First, this work focused on dividing the tweets into 2 phases, before and during COVID-19. In the future, researchers can pursue other methods of structuring the analysis timeline. Second, this study dealt with only the structured textual content of tweets. It would be interesting to also incorporate the presence of image attributes in future studies. Third, as the scope of this study was limited to health care organizations, we did not account for public demographics. Understanding the demographic background of the public engaging with this content is another area that can be explored in future studies. Finally, even though I provide recommendations on how content could be structured on Twitter for increasing user engagement, organizations should consider their priorities and judgment, including ensuring they combat misinformation effectively.

3.4.3 Conclusion

This study examined the online activity of US and Canadian health care organizations on Twitter. The NLP-based analysis of social media presented here can be incorporated to gauge engagement on the previously published tweets and to generate tweets that create an impact on people accessing health information via SMPs. As organizations continue to leverage SMPs by providing the latest information to the community, predicting a tweet's sentiment before publishing can boost an organization's perception by the public. In conclusion, we found that performing content analysis and sentiment forecasting on an organization's social media usage provides a comprehensive view of how it resonates with society.

Chapter 4

Exploring How Healthcare Organizations Use Twitter: A Discourse Analysis

All of this chapter has been accepted in the International Conference on Intelligent Biology and Medicine (ICIBM 2023) as the following peer-reviewed article :
• Singhal, A., & Mago, V. Exploring How Healthcare Organizations Use Twitter: A Discourse Analysis.

Using the advancements in the field of Natural Language Processing, this chapter provides insights into NLP for health literacy and presents a way for organizations to structure their future content to ensure maximum public engagement.

Keywords : Twitter, causality inference, association rule mining, healthcare organizations, topic modeling

4.1 Introduction

4.1.1 Background and Literature Review

The use of social media platforms for information dissemination has grown significantly in the last decade. Among them, Twitter has emerged as a preferred platform, with 7 out of 10 American adults using it as a daily source of news [136]. Health-related content is one of the crucial types of information shared on the micro-blogging platform, often disseminated by over 2,000 healthcare professionals worldwide [157]. Existing research has focused on analyzing important social dimensions, such as audience reach (e.g., followers and subscribers) and post interactivity (e.g., retweets and likes), to identify the impact of online content [14, 15, 221]. However, there is a dearth of studies that explore the underlying textual patterns in the content shared by healthcare organizations on Twitter.

Twitter has been widely used for real-time infoveillance of health messages and has been studied for content analysis by health practitioners, researchers, and computer scientists [36, 133]. Previous research has examined Twitter data to identify top technologies in the health domain using hashtag analysis [73]. In addition, Broniatowski et al. have explored the use of Twitter bots to amplify vaccine hesitancy [24]. Other studies have focused on health literacy promotion through Twitter and identifying health-related causalities from tweets, such as stress, insomnia, and headache [49, 220]. Twitter has also been investigated for COVID-19-related health beliefs [202]. Various techniques are available to identify stakeholder characteristics on social media platforms (SMPs), including topic-clustering and content analysis [119, 194]. In the context of pharmaceutical companies, research has focused on public opinions related to COVID-19 vaccines [32, 57, 159]. Using techniques such as latent Dirichlet allocation (LDA) and valence-aware dictionary and sentiment reasoner (VADER), researchers have analyzed topics, trends, and sentiments over time [32].

Association rule mining is a data mining technique used to identify relationships between variables in a large dataset by analyzing their co-occurrences. Previous research in social media analysis has utilized association rule mining to understand human behavior [160]. One study on social media involved conducting feature-based opinion analysis using an evolutionary approach with association rule mining, as well as performing interest mining to reveal the relationship between interests and their application value [129].

Other studies have explored the importance of public engagement on Twitter, association rule mining for topic extraction on social media platforms, and the use of centrality measures for detecting influential users on social media [3,103,177]. Recent research has also used association rule mining to extract meaningful information from social media data. For example, researchers have used association rule mining to identify depression symptoms on Twitter and to predict drug usage on the platform [122,192]. Gender differences in internet users have also been identified through aggregation-based data mining algorithms [206]. In addition, researchers have explored the use of word embedding techniques for medication usage classification on Twitter and for the classification of tweets based solely on COVID-19 symptoms [66,90]. Overall, association rule mining has proven to be an effective tool for social media analysis and has been used in a variety of contexts to gain insights into human behavior and social trends.

In recent years, the field of causality analysis in social media has garnered significant interest, as social media platforms generate vast amounts of data that have the potential to reveal causal relationships between variables of interest [62]. For instance, natural experiments have been used to assess the causal impact of social media on political participation, while regression discontinuity design has been employed to identify the causal impact of social media on mental health outcomes [34, 65]. These studies illustrate the potential of causal inference techniques to enhance our understanding of the intricate relationships in social media data.

4.1.2 Objective

In the context of the current global pandemic caused by SARS-CoV-2 (COVID-19), it is essential to understand the impact of social media content on users. As a means of communicating medical information, healthcare organizations make use of social media platforms. Identifying underlying text patterns on Twitter can be challenging due to the large volume of data generated on the platform, as well as the unstructured and noisy nature of the data. In addition, the use of hashtags and mentions can make it challenging to distinguish between topics and identify relevant tweets. Finally, the rapidly changing nature of Twitter content means that patterns and trends can emerge and disappear quickly. Therefore, this study focuses on two primary research questions:

Research question 1: What are the significant text patterns that shape the content of tweets by health agencies and pharmaceutical companies in the US and Canada, and how do they compare with the WHO?

Understanding the topics and information that attract Twitter users can help organizations create content that maximizes user engagement. Identifying underlying text patterns in tweets can provide valuable insights into the topics that are most relevant to a given audience. This information can be used by organizations to tailor their content and messaging to better engage their target audience. In addition, analyzing text patterns can help organizations gain a deeper understanding of how they are discussing specific topics, identify key influencers, and track emerging trends. All of this information can help organizations make informed decisions about their social media strategies and ultimately improve their outreach efforts. Moreover, visualizing the inter-relationship between words of interest, i.e., rules and word patterns, can highlight impactful language styles and text patterns. To achieve these goals, we applied topic modeling and association rule mining to our dataset.

Findings: Pharmaceutical companies shared a more diverse range of content compared to other organizations, while COVID-19 was the most commonly discussed topic across all of them. More details are available in Sections 4.3.1 and 4.3.2.

Research question 2: How can we analyze and evaluate the impact of word patterns on the content shared by healthcare organizations on Twitter?

Twitter users interact with the content shared on the micro-blogging platform in various ways, including likes, reshares, replies, and others. Additionally, every word pattern and rule generated through association rule mining has specific metrics linked with it. To evaluate the relationship between the two, we employed two evaluation metrics: Tweet Popularity and Rule Support. Identifying the relationship between tweet popularity and text patterns can help healthcare professionals understand how users engage with content shared by healthcare organizations on Twitter. In order to determine whether a change in one variable, such as the use of association rule, hashtag, or mention leads to changes in tweet popularity, we conducted causality analysis. This information can be useful in understanding the effectiveness of communication strategies and identifying what types

of content resonate with users, which can lead to more effective dissemination of information. To identify potential confounding factors that could impact tweet popularity, we conducted causality analysis on the following hypotheses:

Hypothesis 1: Twitter posts having top hashtags and mentions receive more retweets, i.e. they are more popular.

Hypothesis 2: Twitter posts having popular association rules receive more retweets, i.e. they are more popular.

Findings: Both hypotheses were confirmed, with the presence of popular association rules resulting in a higher probability of increased tweet popularity. More details are available in Section 4.3.3.

The ultimate objective of this study is to provide valuable insights into the use of textual features for structuring online content, thereby enhancing public engagement. This paper is organized as follows: Section 4.2 describes the dataset, content analysis, association rule mining, and causality analysis. Section 4.3 presents the study's findings, followed by a discussion in Section 4.4. Lastly, Section 4.5 concludes the paper.

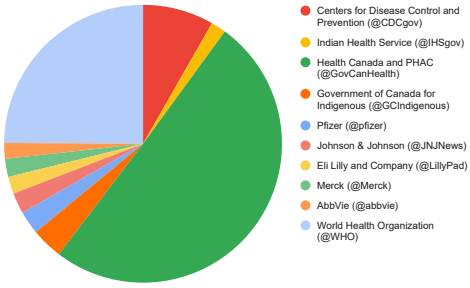
4.2 Materials and Methods

4.2.1 Dataset

This study utilized Twitter data from ten major healthcare organizations in North America and the World Health Organization (WHO). Focusing on a specific geographical region, such as North America, provides us with a context-specific approach, which may affect how health information is communicated and received on social media platforms. This approach enables a deeper understanding of the unique features and factors that impact Twitter usage and the dissemination of health information in that context. Pharmaceutical companies play an important role in developing and producing medical products, including vaccines, which are important in the context of the COVID-19 pandemic. Analyzing their use of Twitter can provide insights into their messaging and communication strategies, which can be useful for both the pharmaceutical companies themselves

Table 4.1: Number of tweets for each organization.

Organization (Twitter Account)	Number of Tweets
Public Health Agencies	
Centers for Disease Control and Prevention (@CDCgov)	8,629
Indian Health Service (@IHSgov)	1,832
Health Canada and PHAC (@GovCanHealth)	52,518
Government of Canada for Indigenous (@GCIndigenous)	3,833
Total	**66,812**
Pharmaceutical Companies	
Pfizer (@pfizer)	2,813
Johnson & Johnson (@JNJNews)	2,538
Eli Lilly and Company (@LillyPad)	2,078
Merck (@Merck)	2,204
AbbVie (@abbvie)	1,913
Total	**11,546**
Non-governmental Organization	
World Health Organization (@WHO)	**25,989**



and for public health organizations looking to work with them. Furthermore, comparing the Twitter usage of health agencies and pharmaceutical companies in the US and Canada with that of the WHO allows us to gain insights into the differences in content and messaging strategies employed by organizations with different levels of reach and influence. This is valuable for organizations looking to improve their own social media strategies, as the WHO's Twitter usage may serve as a benchmark for best practices in health communication on social media. We collected a total of 104,347 tweets from January 01, 2020, to December 31, 2022, using Twitter Academic API for Research v2. National public health agencies, indigenous health agencies, and the top 5 pharmaceutical companies by market capitalization[1] in the United States and Canada were selected for this research. Table 4.1 outlines the number of tweets for each organization, and Fig. 4.1 presents an overview of the research framework.

4.2.2 Content Analysis

In order to understand the textual content of the tweets, we perform the following analyses:

[1]https://www.globaldata.com/companies/top-companies-by-sector/healthcare/us-companies-by-market-cap/

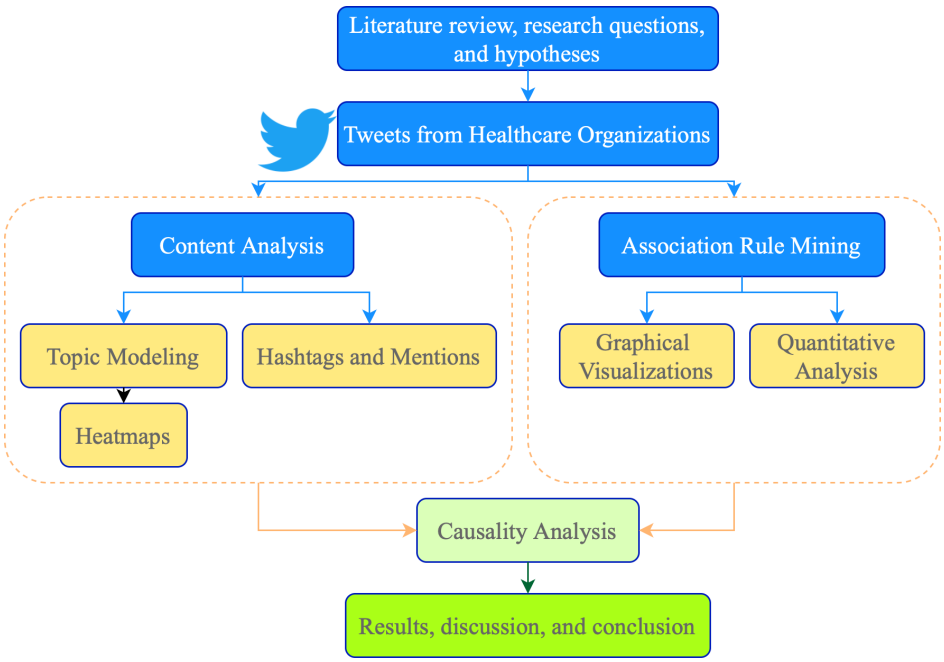


Figure 4.1: Overview of research framework.

Topic Modeling

Topic modeling is a statistical technique widely used in natural language processing to identify latent topics within a collection of documents. In social media, it can be used to analyze the content of tweets shared by healthcare organizations and identify the key themes that emerge. By identifying these themes, we extract the topics that are most relevant to users and can develop content that is tailored to their interests. The process of topic modeling involves identifying a set of latent topics and then analyzing the frequency of words and phrases that are associated with each topic. We first pre-processed the data to remove all non-alphabet characters such as punctuations, numbers, new-line characters, and extra spaces using the regex[2] (regular expression) module 2.2.1. Then, we tokenized it using the nltk[3] 3.2.5 library and performed stemming and lemmatization using PorterStemmer and WordNetLemmatizer, respectively.

In this study, we utilized the term frequency-inverse document frequency (TF-IDF) method to create document embeddings for tweets [115]. The resulting embeddings were then preprocessed and fed into four different clustering algorithms, including la-

[2]https://pypi.org/project/regex/
[3]https://nltk.readthedocs.io/en/latest/

Table 4.2: Model parameters for topic clustering using TF-IDF document embeddings.

Clustering Algorithm	Epochs	Chunk Size	A-priori belief on doc-topic distribution	A-priori belief on topic-word distribution	Gradient descent step size
LDA	50	1000	0.01	0.9	NA
LSI	NA	1000	NA	NA	NA
NMF	50	1000	NA	NA	1
HDP	NA	1000	0.01	NA	1

tent dirichlet allocation (LDA), nonnegative matrix factorization (NMF), latent semantic indexing (LSI), and the hierarchical dirichlet process (HDP). The clustering algorithms were executed five times using varying random seed values to account for the short and noisy nature of tweets. Table 4.2 shows model parameters for topic clustering. To ensure consistency in performance across multiple runs, the coherence scores of the topic models were calculated using both the c_{umass} and c_v methods [144, 164]. The Gensim LDA[4] and Gensim LSI[5] models were used for the analysis, while online NMF[6] for large corpora and online variational inference models were used for NMF and HDP[7] models, respectively.

Heatmaps

Creating heatmaps on topics of tweets shared by healthcare organizations helps us visualize the distribution of topics across different organizations and time periods. This can show which topics are most commonly discussed by healthcare organizations and how these topics may change over time. By examining the heatmap, we identify trends in the distribution of topics and patterns in how different organizations discuss certain topics. This information can be useful for organizations looking to improve their social media strategies, as they can identify which topics are most relevant to their audience and develop content accordingly. Heatmaps also help us identify topics that are commonly discussed by multiple organizations, as these topics may be particularly important or relevant to the broader healthcare community. By identifying these topics, we understand areas of shared interest and potential opportunities for collaboration among healthcare organizations. We created heatmaps to visually analyze the number of tweets for each topic. We used the best-performing topic model to generate them, and each cell in the heatmap represented the total count of tweets for a specific topic by an organization. The

[4]https://radimrehurek.com/gensim/models/ldamodel.html
[5]https://radimrehurek.com/gensim/models/nmf.html
[6]https://radimrehurek.com/gensim/models/lsimodel.html
[7]https://radimrehurek.com/gensim/models/hdpmodel.html

cells in the heatmap are color-coded based on the prevalence of the topic. The darker the color, the higher the prevalence of the topic.

Hashtags and Mentions

Hashtags and mentions can serve as indicators of the conversations and communities that are engaged in a particular topic, as well as the influencers and thought leaders who are driving the discussion. By analyzing their frequency and patterns, we gain a better understanding of how organizations are engaging with health-related topics on social media and which topics are generating the most engagement and interest. This information can be useful for healthcare organizations to tailor their social media strategies and content to better engage with their audience and promote their messaging effectively. The advertools[8] 0.13.2 module was used to analyze the top 10 hashtags and mentions in the data.

4.2.3 Association Rule Mining

Association rule mining is used to uncover the co-occurrence of specific words or phrases, which can provide insights into the topics that are frequently discussed together in tweets. These can be used to inform content creation and messaging strategies for healthcare organizations on social media platforms like Twitter. In addition, association rule mining helps to identify potentially useful combinations of keywords or phrases that can be used to optimize search queries and information retrieval in the context of health information dissemination. By identifying which keywords or phrases are frequently mentioned together in tweets, we can develop more effective search strategies that take into account the associations and relationships between different concepts in the domain of health communication.

The *mlxtend* python library was utilized for association rule mining in our study [9]. The dataset was cleaned (as described in previous section), encoded as Numpy arrays using the *TransactionEncoder()* API and transformed into a one-hot encoded Numpy boolean array using the *fit* and *transform* methods. It was then converted into a *pandas* DataFrame. To perform association rule mining on the tweets, we utilized the Apriori algorithm, which

[8]https://pypi.org/project/advertools/
[9]https://github.com/rasbt/mlxtend

Table 4.3: Grid search parameters used for obtaining association rules.

2*Twitter Group	Support value			Confidence value			Final Support Threshold	Final Confidence Threshold	Number of Rules
	Start	End	Step size	Start	End	Step size			
Public Health Agencies	0.0095	0.105	0.001	0.5	1	0.1	0.1	0.9	980
Pharmaceutical Companies	0.03	0.04	0.001	0.5	1	0.1	0.034	0.5	278
World Health Organization	0.01	0.02	0.001	0.5	1	0.1	0.015	0.8	451

is a classic and widely-used algorithm for mining frequent itemsets [4, 5] by employing the function from *mlxtend.frequent_patterns*. We generated rules of the form X→Y [189] by performing a grid search, where X and Y refer to the antecedent and consequent, respectively. These rules were evaluated based on metrics such as support, confidence, and lift to determine their strength and significance. Table 4.3 displays the parameters that were used for grid search to determine the number of relevant association rules for each Twitter group, and the corresponding parameters utilized for deriving these association rules.

The confidence metric was calculated to determine interesting rules:

$$confidence(A \rightarrow C) = \frac{support(A \rightarrow C)}{support(A)} \;\; ; range \in [0, 1] \tag{4.1}$$

where

$$support(A \rightarrow C) = support(A \cup C) \;\; ; range \in [0, 1] \tag{4.2}$$

For instance, a support threshold of 0.5 (50%) means that a set of items should appear together in at least 50% of all transactions in the database. When both antecedent and consequent always occur together, the confidence is 1. We filter rules using the *lift* metric, which ensures that antecedents and consequents are statistically independent, i.e., $lift \geq 1$.

$$lift(A \rightarrow C) = \frac{confidence(A \rightarrow C)}{support(C)} \;\; ; range \in [0, \infty) \tag{4.3}$$

Graphical Visualizations

Graphical visualizations of association rules obtained provide valuable insights into the relationships between different topics and concepts. By representing these relationships visually, we can more easily identify patterns and trends in the data that might be difficult to discern from raw numbers or statistics alone. In our study, we utilized the D3JS[10] JavaScript library, a popular tool for generating interactive network visualizations, to create hierarchical edge bundling charts. In these charts, we map the antecedents and consequents obtained from the association rules as the source and target, respectively. This helps us to identify clusters of related terms, as well as any outliers or unexpected relationships between terms. Furthermore, graphical visualizations make it easier to communicate the results of the analysis to a broader audience, including healthcare professionals, policymakers, and members of the public who may not have a technical background. The visual nature of these representations can make the findings more accessible and easier to understand, which can be particularly important in the context of public health communication.

Quantitative Analysis

Calculating tweet popularity and rule support for association rules obtained from tweets shared by healthcare organizations is important for evaluating the significance and relevance of the rules. Following Mahdikhani's approach, we measure tweet popularity as the count of retweets [123].

Rule support is a measure of the frequency with which a rule occurs in the data set. By calculating tweet popularity for tweets associated with certain rules, we identify which rules are associated with tweets that have the most engagement, indicating that these topics or themes are more interesting or relevant to the audience. We calculate rule support as the summation of antecedent support, consequent support, overall support, confidence, lift, leverage, and conviction metrics.

[10]https://d3js.org

$$Rule\_support = antecedent\_support + consequent\_support + overall\_support + confidence + lift + leverage + conviction \quad (4.4)$$

4.2.4 Causality Analysis

Causality analysis, also known as causal inference, is a statistical method used to determine if there is a causal relationship between two or more variables [146]. It involves identifying a potential causal relationship between an independent variable and a dependent variable and then determining whether that relationship is actually causal or just a result of some other factor, called a confounding variable. To analyze the causal relationship between the occurrence of top hashtags, mentions, association rules, and tweet popularity, we construct a dataset by comparing their frequencies in the top 10% and the bottom 10% of the tweets ranked by popularity. We assign a binary value of 1 if they occur, and 0 if they do not. We then use the CausalInferenceModel from the CausalNLP[11] package along with an LGBMClassifier that has 500 leaves as the base learner. We consider the overall treatment effect across all observations in the dataset.

4.2.5 Computational Resources

This study utilized the advanced research computing (ARC), research data management (RDM), and research software (RS) resources provided by Compute Canada, now known as the Digital Research Alliance of Canada. Specifically, we used one of the clusters called Graham, which offered the following computing resources:

- Central processing unit (CPU): 2x Intel E5-2683 v4 Broadwell@2.1 GHz
- Memory (RAM): 30 GB

[11] https://github.com/amaiya/causalnlp

Table 4.4: Mean coherence scores for topic modeling using different clustering algorithms.

Clustering Algorithm	Public Health Agencies		Pharmaceutical Companies		World Health Organization	
	c_v	c_{umass}	c_v	c_{umass}	c_v	c_{umass}
LDA	0.4240202647	-4.494327319	0.4937176966	-4.736215923	0.383932226	-4.571343897
NMF	0.5105442417	[HTML]32CB00-3.58084239	0.5955689283	-4.569011046	0.4195732471	[HTML]32CB00-4.223622355
LSI	0.4274006726	-4.217779619	0.450525174	-4.790466135	0.3155738006	-4.245038402
HDP	0.6681490255	-18.13566144	[HTML]32CB000.7406355945	-19.83139046	0.7215923057	-19.2754207

Table 4.5: List of topics obtained for each Twitter group.

C¿X

Public Health Agencies		Pharmaceutical Companies		World Health Organization	
Topic	**Topic Keywords**	**Topic**	**Topic Keywords**	**Topic**	**Topic Keywords**
Communication	['receive','inform', 'reply','offer']	Communication	['shortage','misinformation']	Leadership	['drtedro','meet','report', 'remark']
COVID-19	['covid','pandemic', 'death','vaccine','coronavirus']	COVID-19	['covid','omicron','vaccine', 'virus','coronavirus']	COVID-19	['covid','pandemic','death', 'vaccine','coronavirus']
Community Healthcare	['support','resource', 'family','opportunity','help']	Community Healthcare	['cancer','heart','pregnancy', 'myeloma','gene','haemophilia']	Community Healthcare	['support','people','live', 'protect','safe','risk','care']
General health	['disease','mental', 'health','stigma']	Health announcements	['market','field','campaign']	General Health	['health','disease', 'emergency']
Youth health	['youth','active','profession']	World Regions	['europe','usa','canada']	World Regions	['europe','afro','africa', 'country','countries']

4.3 Results

4.3.1 Content Analysis

The output of topic modeling is a set of topics, each of which is characterized by a set of words or phrases that are most closely associated with that topic. These topics can then be used to discern the content of tweets shared by healthcare organizations and identify patterns and trends in how health information is communicated on social media. The most relevant topic contents were generated by: NMF for public health agencies (c_{umass} = -3.58) and WHO (c_{umass} = -4.22), and HDP for pharmaceutical companies (c_v = 0.74) as shown in Table 4.4. After extracting the contents, we analyzed the topic keywords to suggest a relevant topic name for each, as shown in Table 4.5. This provides an overview of the main topics discussed by each group on Twitter, based on the keywords used in their tweets. For example, the table shows that the Public Health Agencies focus heavily on topics related to COVID-19, such as pandemic, death, vaccine, and coronavirus. On the other hand, the Pharmaceutical Companies group discussed a wider range of topics, such as communication, shortage, and misinformation, in addition to COVID-19-related topics. Similarly, the WHO discussed topics such as world regions, diseases, and health emergencies. Overall, this table provides a perspective into the different priorities and focuses of each Twitter group, which is useful in understanding their messaging and strategies on social media.

Table 4.6: Heatmaps showing topic distribution for each organization.

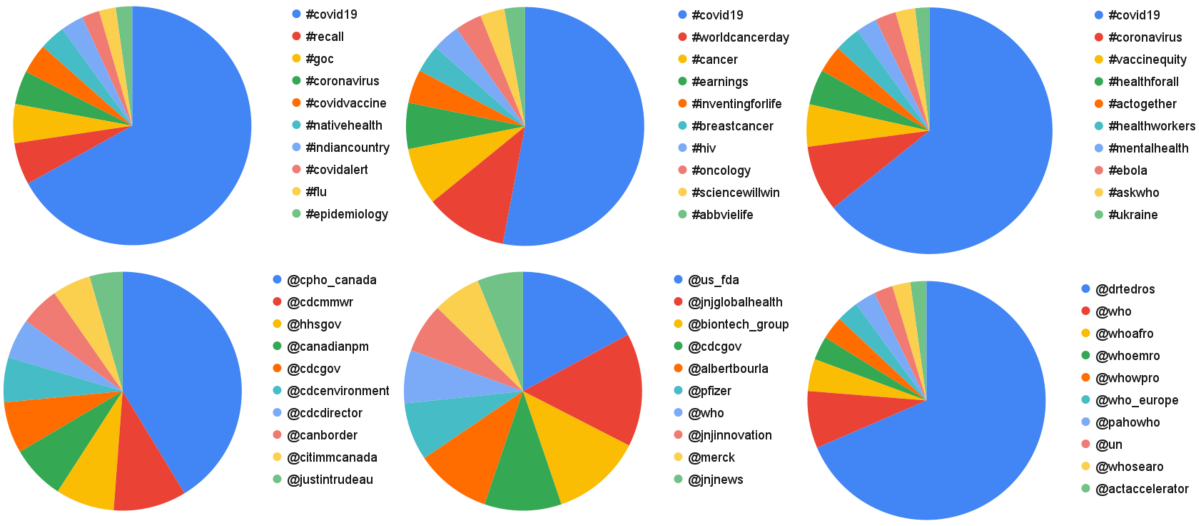
	Topics				
Organization	Communication	COVID-19	Community Health	General Health	Youth Health
CDCgov	[HTML]CCE0F3719	[HTML]C4DBF02306	[HTML]C5DCF02067	[HTML]C6DDF11850	[HTML]CFE2F3195
IHSgov	[HTML]CFE2F3148	[HTML]CEE2F3329	[HTML]CEE1F3438	[HTML]CDE1F3602	[HTML]CFE2F3103
GovCanHealth	[HTML]72A7D717989	[HTML]71A6D618170	[HTML]3D85C627937	[HTML]8AB6DE13353	[HTML]BAD5ED4112
GCIndigenous	[HTML]CEE2F3338	[HTML]CEE2F3381	[HTML]CBE0F2956	[HTML]CEE2F3376	[HTML]CFE2F3156
	Topics				
Organization	Communication	COVID-19	Community Health	Health announcements	World Regions
pfizer	[HTML]CEE1F38	[HTML]3D85C6551	[HTML]5897CF450	[HTML]CCE0F217	[HTML]C9DEF127
JNJNews	[HTML]CFE2F33	[HTML]68A0D3393	[HTML]A5C7E6162	[HTML]C5DCF042	[HTML]CADFF223
Merck	[HTML]CFE2F33	[HTML]A9CAE8145	[HTML]5897CF450	[HTML]CCE0F314	[HTML]CEE1F38
LillyPad	[HTML]CEE1F38	[HTML]C6DDF136	[HTML]BDD6EE73	[HTML]CFE2F35	[HTML]CCE0F314
abbvie	[HTML]CFE2F32	[HTML]BDD7EE70	[HTML]9EC3E4188	[HTML]CADFF223	[HTML]CDE1F313
	Topics				
Organization	Leadership	COVID-19	Community Health	General Health	World Regions
WHO	[HTML]CFE2F31646	[HTML]7EAEDA7165	[HTML]3D85C611486	[HTML]6AA2D48473	[HTML]ABCBE84097

The heatmap in Table 4.6 displays the distribution of topics among the different healthcare organizations. It reveals that GovCanHealth was the most active health agency across all topics. In contrast, the distribution of topics among pharmaceutical companies was more evenly spread, with Pfizer having the highest number of posts related to COVID-19, while Merck, LillyPad, and AbbVie focused more on community health, in line with WHO's activity. The heatmap helps to identify any patterns or trends in the social media activity of the organizations and provides insights into their communication strategies.

The hashtag #covid19 was most frequently used by pharmaceutical companies, public health agencies, and WHO, as shown in Figure 4.2. The most tagged Twitter accounts were the US FDA, the Chief Public Health Officer of Canada, and the Director-General of the World Health Organization, respectively. This provides insight into which individuals and organizations are most active in the healthcare industry on social media platforms, indicating their influence or authority in the field.

4.3.2 Association Rule Mining

The figures in Figure 4.3 display the visualized association rules for each Twitter group. The antecedents (or sources) are represented in blue, while the consequents (or targets) are shown in red. These visualizations illustrate the most frequent association rules present in our data set. Upon analysis, we observed that the association rule pairs from public health agencies and WHO were fewer and more precise as compared to those from pharmaceutical companies. In the case of public health agencies, the most impactful



(a) Public Health Agencies (b) Pharmaceutical Companies (c) World Health Organization

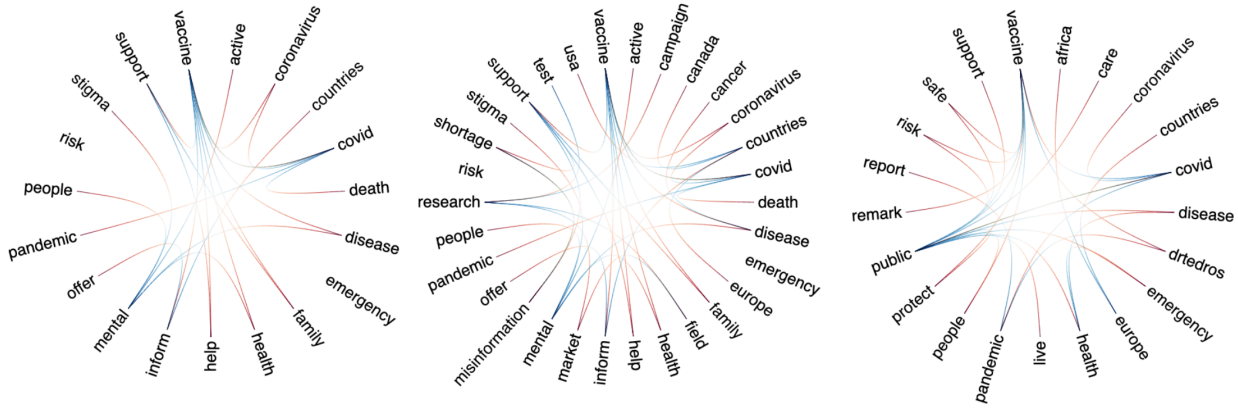
Figure 4.2: Top hashtags and mentions for each group of healthcare organizations.

Table 4.7: Top association rules and performance metrics obtained.

Twitter Group	Antecedents	Consequents	Antecedent support	[HTML]FFFFFF Consequent support	[HTML]FFFFFF Overall support	Confidence	Lift	Leverage	Conviction
Public Health Agencies	covid	vaccine	0.11	0.11	0.10	0.99	8.61	0.10	6851.90
Pharmaceutical Companies	test	research	0.03	0.03	0.03	0.99	28.72	0.03	388.03
World Health Organization	public	health	0.038	0.23	0.03	0.80	3.42	0.02	3.85

antecedent-consequent pairs were associated with COVID-19, including 'covid-vaccine' and 'vaccine-mental'. Conversely, the highest ranked association rules from pharmaceutical companies such as 'test-research', 'market-research', and 'vaccine-covid' explored topics beyond the pandemic, such as communication and innovation. Association rules obtained for WHO included 'public-health' in addition to rules denoting regional WHO offices such as 'europe-africa' (for WHO Europe and Africa). These findings can inform organizations in the healthcare industry on how to structure their tweets to achieve maximum engagement from their target audience.

We rank tweets in each Twitter group based on their Tweet Popularity metric and association rules according to Rule Support in descending order. Table 4.7 lists the top association rules and performance metrics obtained for each Twitter group, which are a combination of individual words.



(a) Public Health Agencies (b) Pharmaceutical Companies (c) World Health Organization

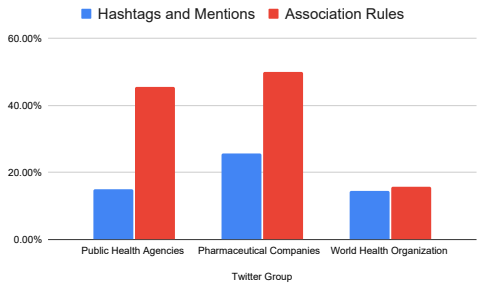
Figure 4.3: Graph networks showing Antecedent - Consequent pairs. Public health agencies and WHO generate sparse graphs focused on COVID-19, while pharmaceutical companies generate a denser graph with words from different topics.

4.3.3 Causality Analysis

Table 4.8 summarizes the findings of our study that examined the effect of hashtags, mentions, and association rules on the popularity of tweets in the online sphere. The analysis revealed that tweets that included these elements were more likely to be shared. The study tested two hypotheses. The first hypothesis showed that the presence of top hashtags and mentions increased the likelihood of a post being shared by 14.90% and 14.55% for public health agencies and WHO, respectively. In contrast, for pharmaceutical companies, the probability increased significantly by 25.70%. The second hypothesis examined the impact of top association rules and found that their presence increased the probability of a tweet being popular by 45.05% and 50.05% for public health agencies and pharmaceutical companies, respectively. However, the chance of popularity for WHO was lower at 15.70%, potentially due to its global presence and the higher impact of its regional arms, as suggested by the topic modeling results. Overall, this highlights the importance of using association rules as compared to hashtags and mentions to increase the likelihood of a tweet being shared in the online sphere.

Table 4.8: Results of causality analysis using two hypotheses to analyze the impact on tweet popularity.

Twitter Group	Hypothesis 1: Increase in Tweet popularity using Hashtags and Mentions	Hypothesis 2: Increase in Tweet popularity Association Rules
Public Health Agencies	[HTML]FFFFFF14.90%	[HTML]FFFFFF45.50%
Pharmaceutical Companies	[HTML]FFFFFF25.70%	[HTML]FFFFFF50.05%
World Health Organization	[HTML]FFFFFF14.55%	[HTML]FFFFFF15.70%



4.4 Discussion

In this study, we performed content analysis, association rule mining, and causality inference on a large database of tweets from healthcare organizations to understand the textual patterns and their impact on driving engagement. Based on our analyses, the principal findings are:

- RQ1: What are the significant text patterns that shape the content of tweets by health agencies and pharmaceutical companies in the US and Canada, and how do they compare with the WHO?

  The study used topic modeling to identify the main text patterns present in tweets by health agencies and pharmaceutical companies in the US and Canada, as well as the WHO. The analysis revealed that public health agencies and the WHO focused heavily on COVID-19-related topics, while pharmaceutical companies covered a wider range of topics, including communication and innovation. The distribution of topics among the different organizations was also visualized using a heatmap, which showed that GovCanHealth was the most active health agency across all topics. The study also identified the most frequently used hashtag (#covid19) and tagged Twitter accounts in the healthcare industry, providing insight into the most active and influential individuals and organizations. Finally, the study analyzed association rules to identify the most impactful antecedent-consequent pairs in tweets by each group. The findings suggest that public health agencies and the WHO generated fewer but more precise association rules related to COVID-19, while pharmaceutical companies explored topics beyond the pandemic. These results can help organizations in the healthcare industry to structure their tweets to achieve maximum

engagement from their target audience, and this approach is especially beneficial for organizations that seek to align their content with a common goal, as it enables them to synergize their efforts toward creating effective messaging.

- RQ2: How can we analyze and evaluate the impact of word patterns on the content shared by healthcare organizations on Twitter?

  In order to effectively analyze the impact of word patterns, we calculated two metrics: tweet popularity (count of retweets) and rule support (sum of all performance metrics). These metrics can be used to rank tweets and association rules in each Twitter group. In addition to analyzing association rules, the study also examined the effect of hashtags, mentions, and association rules on the popularity of tweets in the online sphere. The analysis revealed that tweets that included these elements were more likely to be shared. We also tested two hypotheses, which showed that the presence of top hashtags and mentions increased the likelihood of a post being shared, and that the presence of top association rules significantly increased the probability of a tweet being popular.

  Overall, this study highlights the importance of using association rules as compared to hashtags and mentions to increase the likelihood of a tweet being shared in the online sphere. It also provides insights into the impact of word patterns on the content shared by healthcare organizations on Twitter and offers a way to evaluate their effectiveness. By analyzing the language and style used in popular tweets, organizations can gain insights into what resonates with their audience and adjust their messaging accordingly. This leads to better communication of health information, increased engagement, and better health outcomes. Researchers can provide valuable insights to help organizations improve their communication strategies and better disseminate health information on social media platforms like Twitter.

4.4.1 Limitations and Future Research Directions

This study focuses on textual features of Twitter content and their relationship to user engagement. Although causality analysis is a powerful tool for identifying causal relationships between variables, it is important to recognize that causality cannot be established definitively in all cases. There may be other variables, such as images or videos, that are not included in the analysis that could be driving the observed associations. Future re-

search could also focus on investigating the effectiveness of social media campaigns and interventions on health-related outcomes.

4.5 Conclusions

As social media platforms become ubiquitous in our daily lives, healthcare organizations can leverage them to increase public engagement. This study examined the content shared by healthcare organizations on Twitter by performing content analysis, association rule mining and causality analysis. NLP methods, such as topic modeling, help identify the overall themes and topics of the tweets, but association rule mining can help identify which words, phrases, or language patterns are associated with higher or lower tweet popularity, allowing organizations to adjust their messaging and communication strategies accordingly. Using popular association rules also significantly increases the probability of a tweet getting reshared across all categories. Overall, the methodology presented here can help healthcare organizations fine-tune their content for their audience.

Chapter 5

Conclusion

This thesis contributes to the area of natural language processing for social media and healthcare interventions by providing methodologies for analyzing and increasing public engagement. Additionally, it also demonstrates the potential of FATE-infused algorithms to deliver trustworthy and equitable results. All in all, this thesis proposes innovative and effective NLP strategies catering to a wide global audience in the online sphere while catering to the needs of an individual.

At first, this thesis determined the progress made in the area of FATE of AI in the last ten years through a systematic review of scholarly research articles. Through a well-developed search strategy that selected top-cited publications, the most prominent solutions to FATE are identified and compared in terms of computational methods, approaches, and evaluation metrics allowing for a thorough overview of the field's development.

Then, my study on sentiment and engagement analysis using CardiffNLP's twitter-roberta- base-sentiment model contributes to how social media usage by public health agencies, nongovernment organizations (NGOs), and pharmaceutical companies resonates with society. The research includes a study of tweets' sentiments using 16 univariate forecasting models, to effectively present the model topics and best-performing sentiment-forecasting models.

Then, shifting to a newer aspect of discourse analysis, Chapter 4 demonstrated the potential of causality analysis to identify confounding factors that shape the text patterns resulting in impactful tweet content. Overall, using popular association rules helps an organization come ahead of its competitors in the online sphere. This study underscores the pivotal role of natural language processing techniques in advancing health literacy and provides actionable insights for healthcare organizations to optimize their future content strategies for maximal public engagement.

Investigating deeper into causality analysis, using other variables such as images or videos is a promising research area as outlined in this thesis. Understanding the demographic background of the social media users, in addition to the impact of targeted social media campaigns on health-related campaigns can also be explored based on the evidence presented earlier. Another potential direction for future research could be developing and evaluating FATE frameworks and strategies to streamline their use in algorithms for social media and healthcare.

Throughout this thesis, the importance of NLP in improving healthcare outcomes has been underscored. Furthermore, it has emphasized the significance of incorporating FATE technology and design principles into algorithms to ensure fairness and equity for all. By addressing these critical aspects, this research has paved the way for advancements in NLP applications in social media and healthcare, ultimately benefiting individuals and communities alike.

Bibliography

[1] ABUALIGAH, L., ALFAR, H. E., SHEHAB, M., AND HUSSEIN, A. M. A. Sentiment analysis in healthcare: a brief review. Recent advances in NLP: the case of Arabic language (2020), 129–141.

[2] ADADI, A., AND BERRADA, M. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE access 6 (2018), 52138–52160.

[3] AGOUTI, T. Graph-based modeling using association rule mining to detect influential users in social networks. Expert Systems with Applications (2022), 117436.

[4] AGRAWAL, R., IMIELINSKI, T., AND SWAMI, A. Mining associations between sets of items in large databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data (1993), pp. 207–216.

[5] AGRAWAL, R., SRIKANT, R., ET AL. Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB (1994), vol. 1215, Citeseer, pp. 487–499.

[6] AIELLO, A. E., RENSON, A., AND ZIVICH, P. Social media-and internet-based disease surveillance for public health. Annual review of public health 41 (2020), 101.

[7] ALOMARI, K. M., ELSHERIF, H. M., AND SHAALAN, K. Arabic tweets sentimental analysis using machine learning. In Advances in Artificial Intelligence: From Theory to Practice: 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017, Arras, France, June 27-30, 2017, Proceedings, Part I 30 (2017), Springer, pp. 602–610.

[8] AMANN, J., BLASIMME, A., VAYENA, E., FREY, D., MADAI, V. I., AND CONSORTIUM, P. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC medical informatics and decision making 20 (2020), 1–9.

[9] ARNOLD, T., AND KASENBERG, D. Value alignment or misalignment â€“what will keep systems accountable? In AAAI Workshop on AI, Ethics, and Society (2017).

[10] ARRIETA, A. B., DÍAZ-RODRÍGUEZ, N., DEL SER, J., BENNETOT, A., TABIK, S., BARBADO, A., GARCÍA, S., GIL-LÓPEZ, S., MOLINA, D., BENJAMINS, R., ET AL. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion 58 (2020), 82–115.

[11] ASUNCION, A., AND NEWMAN, D. Uci machine learning repository, 2007.

[12] ATTARD-FROST, B., DE LOS RÍOS, A., AND WALTERS, D. R. The ethics of ai business practices: a review of 47 ai ethics guidelines. AI and Ethics (2022), 1–18.

[13] AZEROUAL, O., SAAKE, G., AND SCHALLEHN, E. Analyzing data quality issues in research information systems via data profiling. International Journal of Information Management 41 (2018), 50–56.

[14] BAXI, M. K., PHILIP, J., AND MAGO, V. Resilience of political leaders and healthcare organizations during covid-19. PeerJ Computer Science 8 (2022), e1121.

[15] BAXI, M. K., SHARMA, R., AND MAGO, V. Studying topic engagement and synergy among candidates for 2020 us elections. Social Network Analysis and Mining 12, 1 (2022), 136.

[16] BEAR DON'T WALK IV, O. J., REYES NIEVA, H., LEE, S. S.-J., AND ELHADAD, N. A scoping review of ethics considerations in clinical natural language processing. JAMIA open 5, 2 (2022), ooac039.

[17] BELLAMY, R. K., DEY, K., HIND, M., HOFFMAN, S. C., HOUDE, S., KANNAN, K., LOHIA, P., MARTINO, J., MEHTA, S., MOJSILOVIĆ, A., ET AL. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development 63, 4/5 (2019), 4–1.

[18] BENETOLI, A., CHEN, T., AND ASLANI, P. How patients' use of social media impacts their interactions with healthcare professionals. Patient education and counseling 101, 3 (2018), 439–444.

[19] BERTINO, E., MERRILL, S., NESEN, A., AND UTZ, C. Redefining data transparency: A multidimensional approach. Computer 52, 1 (2019), 16–26.

[20] BHATIA-LIN, A., BOON-DOOLEY, A., ROBERTS, M. K., PRONAI, C., FISHER, D., PARKER, L., ENGSTROM, A., INGRAHAM, L., AND DARNELL, D. Ethical and regulatory considerations for using social media platforms to locate and track research participants. The American Journal of Bioethics 19, 6 (2019), 47–61.

[21] BIRD, S., DUDÍK, M., EDGAR, R., HORN, B., LUTZ, R., MILAN, V., SAMEKI, M., WALLACH, H., AND WALKER, K. Fairlearn: A toolkit for assessing and improving fairness in ai. Microsoft, Tech. Rep. MSR-TR-2020-32 (2020).

[22] BLACKLAWS, C. Algorithms: transparency and accountability. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376, 2128 (2018), 20170351.

[23] BOSE, R., AND FREW, J. Lineage retrieval for scientific data processing: a survey. ACM Computing Surveys (CSUR) 37, 1 (2005), 1–28.

[24] BRONIATOWSKI, D. A., JAMISON, A. M., QI, S., ALKULAIB, L., CHEN, T., BENTON, A., QUINN, S. C., AND DREDZE, M. Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. American journal of public health 108, 10 (2018), 1378–1384.

[25] BRUNDAGE, M., AVIN, S., WANG, J., BELFIELD, H., KRUEGER, G., HADFIELD, G., KHLAAF, H., YANG, J., TONER, H., FONG, R., ET AL. Toward trustworthy ai development: mechanisms for supporting verifiable claims. arXiv preprint arXiv:2004.07213 (2020).

[26] BUCHER, M., HERBIN, S., AND JURIE, F. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14 (2016), Springer, pp. 730–746.

[27] BURGESS, K., HART, D., ELSAYED, A., CERNY, T., BURES, M., AND TISNOVSKY, P. Visualizing architectural evolution via provenance tracking: a systematic review. In Proceedings of the Conference on Research in Adaptive and Convergent Systems (2022), pp. 83–91.

[28] CARVALHO, D. V., PEREIRA, E. M., AND CARDOSO, J. S. Machine learning interpretability: A survey on methods and metrics. Electronics 8, 8 (2019), 832.

[29] CHAKRABORTI, T., PATRA, A., AND NOBLE, J. A. Contrastive fairness in machine learning. IEEE Letters of the Computer Society 3, 2 (2020), 38–41.

[30] CHAKRABORTY, S., TOMSETT, R., RAGHAVENDRA, R., HARBORNE, D., ALZANTOT, M., CERUTTI, F., SRIVASTAVA, M., PREECE, A., JULIER, S., RAO, R. M., ET AL. Interpretability of deep learning models: A survey of results. In 2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI) (2017), IEEE, pp. 1–6.

[31] CHANDRASEKARAN, D., AND MAGO, V. Evolution of semantic similarity—a survey. ACM Computing Surveys (CSUR) 54, 2 (2021), 1–37.

[32] CHANDRASEKARAN, R., MEHTA, V., VALKUNDE, T., AND MOUSTAKAS, E. Topics, trends, and sentiments of tweets about the covid-19 pandemic: Temporal infoveillance study. Journal of medical Internet research 22, 10 (2020), e22624.

[33] CHEN, J., AND WANG, Y. Social media use for health purposes: systematic review. Journal of medical Internet research 23, 5 (2021), e17917.

[34] CHEN, S., GELDSETZER, P., AND BÄRNIGHAUSEN, T. The causal effect of retirement on stress in older adults in china: A regression discontinuity study. SSM-population health 10 (2020), 100462.

[35] CHOULDECHOVA, A., AND ROTH, A. A snapshot of the frontiers of fairness in machine learning. Communications of the ACM 63, 5 (2020), 82–89.

[36] COLDITZ, J. B., CHU, K.-H., EMERY, S. L., LARKIN, C. R., JAMES, A. E., WELLING, J., AND PRIMACK, B. A. Toward real-time infoveillance of twitter health messages. American journal of public health 108, 8 (2018), 1009–1014.

[37] CONWAY, M., AND O'CONNOR, D. Social media, big data, and mental health: current advances and ethical implications. Current opinion in psychology 9 (2016), 77–82.

[38] COURTNEY, K., SHABESTARI, O., AND KUO, A. M.-H. The use of social media in healthcare: organizational, clinical, and patient perspectives. Enabling health and healthcare through ICT: available, tailored and closer 183 (2013), 244.

[39] COVID, C., TEAM, R., JORDEN, M. A., RUDMAN, S. L., VILLARINO, E., HOFERKA, S., PATEL, M. T., BEMIS, K., SIMMONS, C. R., JESPERSEN, M., ET AL. Evidence for limited early spread of covid-19 within the united states, january–february 2020. Morbidity and Mortality Weekly Report 69, 22 (2020), 680.

[40] CRAWLEY, A. W., DIVI, N., AND SMOLINSKI, M. S. Using timeliness metrics to track progress and identify gaps in disease surveillance. Health security 19, 3 (2021), 309–317.

[41] CROCKETT, M. J. Models of morality. Trends in cognitive sciences 17, 8 (2013), 363–366.

[42] DANILUK, M., DABROWSKI, J., RYCHALSKA, B., AND GOLUCHOWSKI, K. Synerise at recsys 2021: Twitter user engagement prediction with a fast neural model. In Proceedings of the Recommender Systems Challenge 2021. 2021, pp. 15–21.

[43] DATTA, A., SEN, S., AND ZICK, Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In 2016 IEEE symposium on security and privacy (SP) (2016), IEEE, pp. 598–617.

[44] DAVIS, K. Ethics of Big Data: Balancing risk and innovation. " O'Reilly Media, Inc.", 2012.

[45] DEEPA, N., PRABADEVI, B., MADDIKUNTA, P. K., GADEKALLU, T. R., BAKER, T., KHAN, M. A., AND TARIQ, U. An ai-based intelligent system for healthcare analysis using ridge-adaline stochastic gradient descent classifier. The Journal of Supercomputing 77 (2021), 1998–2017.

[46] DENECKE, K., AND NEJDL, W. How valuable is medical social media data? content analysis of the medical web. Information Sciences 179, 12 (2009), 1870–1880.

[47] DIAKOPOULOS, N., AND KOLISKA, M. Algorithmic transparency in the news media. Digital journalism 5, 7 (2017), 809–828.

[48] DIXON, L., LI, J., SORENSEN, J., THAIN, N., AND VASSERMAN, L. Measuring and mitigating unintended bias in text classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (2018), pp. 67–73.

[49] DOAN, S., YANG, E. W., TILAK, S. S., LI, P. W., ZISOOK, D. S., AND TORII, M. Extracting health-related causality from twitter messages using natural language processing. BMC medical informatics and decision making 19, 3 (2019), 71–77.

[50] DRABIAK, K., AND WOLFSON, J. What should health care organizations do to reduce billing fraud and abuse? AMA Journal of Ethics 22, 3 (2020), 221–231.

[51] DUA, D., GRAFF, C., ET AL. Uci machine learning repository, 2017. URL http://archive. ics. uci. edu/ml 7, 1 (2017).

[52] DUBBERLEY, S., KOENIG, A., AND MURRAY, D. Digital witness: using open source information for human rights investigation, documentation, and accountability. Oxford University Press, USA, 2020.

[53] DUBEY, A. D. Twitter sentiment analysis during covid-19 outbreak. Available at SSRN 3572023 (2020).

[54] ENARSSON, T., ENQVIST, L., AND NAARTTIJÄRVI, M. Approaching the human in the loop–legal perspectives on hybrid human/algorithmic decision-making in three contexts. Information & Communications Technology Law 31, 1 (2022), 123–153.

[55] FAN, C.-Y., CHANG, P.-C., LIN, J.-J., AND HSIEH, J. A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Applied Soft Computing 11, 1 (2011), 632–644.

[56] FERRO, D. B., BRAILSFORD, S., BRAVO, C., AND SMITH, H. Improving healthcare access management by predicting patient no-show behaviour. Decision Support Systems 138 (2020), 113398.

[57] FISHER, A., PATEL, N., PATEL, P., PATEL, P., KRISHNANKUTTY, V., BHAT, V., VALANI, P., MAGO, V., AND RAO, A. An ethical visualization of the northcovid-19 model. PeerJ Computer Science 8 (2022), e980.

[58] GAGNON, K., AND SABUS, C. Professionalism in a digital age: opportunities and considerations for using social media in health care. Physical therapy 95, 3 (2015), 406–414.

[59] GAO, S., HE, L., CHEN, Y., LI, D., LAI, K., ET AL. Public perception of artificial intelligence in medical care: content analysis of social media. Journal of Medical Internet Research 22, 7 (2020), e16649.

[60] GARG, A., AND MAGO, V. Role of machine learning in medical research: A survey. Computer science review 40 (2021), 100370.

[61] GARG, M., SAXENA, C., KRISHNAN, V., JOSHI, R., SAHA, S., MAGO, V., AND DORR, B. J. Cams: an annotated corpus for causal analysis of mental health issues in social media posts. arXiv preprint arXiv:2207.04674 (2022).

[62] GEORGE, G., OSINGA, E. C., LAVIE, D., AND SCOTT, B. A. Big data and data science methods for management research, 2016.

[63] GHASSAMI, A., KHODADADIAN, S., AND KIYAVASH, N. Fairness in supervised learning: An information theoretic approach. In 2018 IEEE international symposium on information theory (ISIT) (2018), IEEE, pp. 176–180.

[64] GHOSH, A., GENUIT, L., AND REAGAN, M. Characterizing intersectional group fairness with worst-case comparisons. In Artificial Intelligence Diversity, Belonging, Equity, and Inclusion (2021), PMLR, pp. 22–34.

[65] GIL DE ZÚÑIGA, H., MOLYNEUX, L., AND ZHENG, P. Social media, political expression, and political participation: Panel analysis of lagged and concurrent relationships. Journal of communication 64, 4 (2014), 612–634.

[66] GILBERT, J.-P., NIU, J., DE MONTIGNY, S., NG, V., AND REES, E. Machine learning identification of self-reported covid-19 symptoms from tweets in canada. In International Workshop on Health Intelligence (2021), Springer, pp. 101–111.

[67] GILPIN, L. H., BAU, D., YUAN, B. Z., BAJWA, A., SPECTER, M., AND KAGAL, L. Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA) (2018), IEEE, pp. 80–89.

[68] GOLDER, S., AHMED, S., NORMAN, G., AND BOOTH, A. Attitudes toward the ethics of research using social media: a systematic review. Journal of medical internet research 19, 6 (2017), e195.

[69] GOLUBEV, A., AND LOUKACHEVITCH, N. Improving results on russian sentiment datasets. In Artificial Intelligence and Natural Language: 9th Conference, AINL 2020, Helsinki, Finland, October 7–9, 2020, Proceedings 9 (2020), Springer, pp. 109–121.

[70] GÓMEZ-GONZÁLEZ, E., GOMEZ, E., MÁRQUEZ-RIVAS, J., GUERRERO-CLARO, M., FERNÁNDEZ-LIZARANZU, I., RELIMPIO-LÓPEZ, M. I., DORADO, M. E., MAYORGA-BUIZA, M. J., IZQUIERDO-AYUSO, G., AND CAPITÁN-MORALES, L. Artificial intelligence in medicine and healthcare: a review and classification of current and near-future applications and their ethical and social impact. arXiv preprint arXiv:2001.09778 (2020).

[71] GRAJALES III, F. J., SHEPS, S., HO, K., NOVAK-LAUSCHER, H., AND EYSENBACH, G. Social media: a review and tutorial of applications in medicine and health care. Journal of medical Internet research 16, 2 (2014), e2912.

[72] GRANDINI, M., BAGLI, E., AND VISANI, G. Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756 (2020).

[73] GROVER, P., KAR, A. K., AND DAVIES, G. "technology enabled health"–insights from twitter analytics with a socio-technical perspective. International Journal of Information Management 43 (2018), 85–97.

[74] GU, D., GAO, Y., CHEN, K., SHI, J., LI, Y., AND CAO, Y. Electricity theft detection in ami with low false positive rate based on deep learning and evolutionary algorithm. IEEE Transactions on Power Systems 37, 6 (2022), 4568–4578.

[75] GUTTMAN, N. Ethical issues in health promotion and communication interventions. In Oxford research encyclopedia of communication. 2017.

[76] HAGERTY, A., AND RUBINOV, I. Global ai ethics: a review of the social impacts and ethical implications of artificial intelligence. arXiv preprint arXiv:1907.07892 (2019).

[77] HAMAN, M. The use of twitter by state leaders and its impact on the public during the covid-19 pandemic. Heliyon 6, 11 (2020), e05540.

[78] HANSEN, E. Hipaa (health insurance portability and accountability act) rules: federal and state enforcement. Medical Interface 10, 8 (1997), 96–8.

[79] HARPER, R., AND SOUTHERN, J. A bayesian deep learning framework for end-to-end prediction of emotion from heartbeat. IEEE transactions on affective computing 13, 2 (2020), 985–991.

[80] HE, J., BAXTER, S. L., XU, J., XU, J., ZHOU, X., AND ZHANG, K. The practical implementation of artificial intelligence technologies in medicine. Nature medicine 25, 1 (2019), 30–36.

[81] HERTWECK, C., HEITZ, C., AND LOI, M. On the moral justification of statistical parity. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021), pp. 747–757.

[82] HOSSIN, M., SULAIMAN, M., MUSTAPHA, A., MUSTAPHA, N., AND RAHMAT, R. A hybrid evaluation metric for optimizing classifier. In 2011 3rd Conference on Data Mining and Optimization (DMO) (2011), IEEE, pp. 165–170.

[83] HUSSAIN, A., TAHIR, A., HUSSAIN, Z., SHEIKH, Z., GOGATE, M., DASHTIPOUR, K., ALI, A., AND SHEIKH, A. Artificial intelligence–enabled analysis of public attitudes on facebook and twitter toward covid-19 vaccines in the united kingdom and the united states: Observational study. Journal of medical Internet research 23, 4 (2021), e26627.

[84] HUTCHINSON, B., SMART, A., HANNA, A., DENTON, E., GREER, C., KJARTANSSON, O., BARNES, P., AND MITCHELL, M. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021), pp. 560–575.

[85] IYER, R., LI, Y., LI, H., LEWIS, M., SUNDAR, R., AND SYCARA, K. Transparency and explanation in deep reinforcement learning neural networks. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (2018), pp. 144–150.

[86] JADON, S. A survey of loss functions for semantic segmentation. In 2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB) (2020), IEEE, pp. 1–7.

[87] JAKESCH, M., BUÇINCA, Z., AMERSHI, S., AND OLTEANU, A. How different groups prioritize ethical values for responsible ai. In 2022 ACM Conference on Fairness, Accountability, and Transparency (2022), pp. 310–323.

[88] JANG, H., REMPEL, E., ROTH, D., CARENINI, G., AND JANJUA, N. Z. Tracking covid-19 discourse on twitter in north america: Infodemiology study using topic modeling and aspect-based sentiment analysis. Journal of medical Internet research 23, 2 (2021), e25431.

[89] JANSSEN, M., HARTOG, M., MATHEUS, R., YI DING, A., AND KUK, G. Will algorithms blind people? the effect of explainable ai and decision-makers' experience on ai-supported decision-making in government. Social Science Computer Review 40, 2 (2022), 478–493.

[90] JIANG, K., FENG, S., CALIX, R. A., AND BERNARD, G. R. Assessment of word embedding techniques for identification of personal experience tweets pertaining to medication uses. In International Workshop on Health Intelligence (2019), Springer, pp. 45–55.

[91] JOHNSON, S. L. Ai, machine learning, and ethics in health care. Journal of Legal Medicine 39, 4 (2019), 427–441.

[92] KALKMAN, S., MOSTERT, M., GERLINGER, C., VAN DELDEN, J. J., AND VAN THIEL, G. J. Responsible data sharing in international health research: a systematic review of principles and norms. BMC medical ethics 20 (2019), 1–13.

[93] KAPLAN, R. M., BURSTEIN, J., HARPER, M., AND PENN, G. Human language technologies: the 2010 annual conference of the north american chapter of the association for computational linguistics. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (2010).

[94] KASS, N. E., AND FADEN, R. R. Ethics and learning health care: the essential roles of engagement, transparency, and accountability. Learning Health Systems 2, 4 (2018), e10066.

[95] KAUR, D., USLU, S., DURRESI, A., BADVE, S., AND DUNDAR, M. Trustworthy explainability acceptance: A new metric to measure the trustworthiness of interpretable ai medical diagnostic systems. In Complex, Intelligent and Software Intensive Systems: Proceedings of the 15th International Conference on Complex, Intelligent and Software Intensive Systems (CISIS-2021) (2021), Springer, pp. 35–46.

[96] KAZIM, E., AND KOSHIYAMA, A. S. A high-level overview of ai ethics. Patterns 2, 9 (2021), 100314.

[97] KERIKMÄE, T., AND PÄRN-LEE, E. Legal dilemmas of estonian artificial intelligence strategy: in between of e-society and global race. Ai & Society 36 (2021), 561–572.

[98] KESHK, M., MOUSTAFA, N., SITNIKOVA, E., AND TURNBULL, B. Privacy-preserving big data analytics for cyber-physical systems. Wireless Networks (2022), 1–9.

[99] KING, J. The instrumental value of legal accountability. Oxford University Press, 2013.

[100] KINGTON, R. S., ARNESEN, S., CHOU, W.-Y. S., CURRY, S. J., LAZER, D., AND VILLARRUEL, A. M. Identifying credible sources of health information in social media: Principles and attributes. NAM perspectives 2021 (2021).

[101] KO, R. K., KIRCHBERG, M., AND LEE, B. S. From system-centric to data-centric logging-accountability, trust & security in cloud computing. In 2011 Defense Science Research Conference and Expo (DSR) (2011), IEEE, pp. 1–4.

[102] KOFOD-PETERSEN, A. How to do a structured literature review in computer science. Ver. 0.1 1 (2012).

[103] KOUKARAS, P., TJORTJIS, C., AND ROUSIDIS, D. Mining association rules from covid-19 related twitter data to discover word patterns, topics and inferences. Information Systems 109 (2022), 102054.

[104] KOUMPOUROS, Y., TOULIAS, T. L., AND KOUMPOUROS, N. The importance of patient engagement and the use of social media marketing in healthcare. Technology and Health Care 23, 4 (2015), 495–507.

[105] KYNKÄÄNNIEMI, T., KARRAS, T., LAINE, S., LEHTINEN, J., AND AILA, T. Improved precision and recall metric for assessing generative models. Advances in Neural Information Processing Systems 32 (2019).

[106] LAGIOIA, F., ROVATTI, R., AND SARTOR, G. Algorithmic fairness through group parities? the case of compas-sapmoc. AI & SOCIETY (2022), 1–20.

[107] LATONERO, M. Governing artificial intelligence: Upholding human rights & dignity.

[108] LEIDNER, J. L., AND PLACHOURAS, V. Ethical by design: Ethics best practices for natural language processing. In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing (2017), pp. 30–40.

[109] LEIKAS, J., KOIVISTO, R., AND GOTCHEVA, N. Ethical framework for designing autonomous intelligent systems. Journal of Open Innovation: Technology, Market, and Complexity 5, 1 (2019), 18.

[110] LEONELLI, S., LOVELL, R., WHEELER, B. W., FLEMING, L., AND WILLIAMS, H. From fair data to fair data use: Methodological data fairness in health-related social media research. Big Data & Society 8, 1 (2021), 20539517211010310.

[111] LESLIE, D. Understanding artificial intelligence ethics and safety. arXiv preprint arXiv:1906.05684 (2019).

[112] LI, H., AND SAKAMOTO, Y. Social impacts in social media: An examination of perceived truthfulness and sharing of information. Computers in Human Behavior 41 (2014), 278–287.

[113] LI, Q., AND LI, Q. Overview of data visualization. Embodying Data: Chinese Aesthetics, Interactive Visualization and Gaming Technologies (2020), 17–47.

[114] LI, Y., VINZAMURI, B., AND REDDY, C. K. Constrained elastic net based knowledge transfer for healthcare information exchange. Data Mining and Knowledge Discovery 29 (2015), 1094–1112.

[115] LILLEBERG, J., ZHU, Y., AND ZHANG, Y. Support vector machines and word2vec for text classification with semantic features. In 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC) (2015), IEEE, pp. 136–140.

[116] LINK, A. N., AND SCOTT, J. T. Public accountability: Evaluating technology-based institutions. Springer Science & Business Media, 2012.

[117] LIPTON, Z. C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16, 3 (2018), 31–57.

[118] LIVINGSTON, J. D., MILNE, T., FANG, M. L., AND AMARI, E. The effectiveness of interventions for reducing stigma related to substance use disorders: a systematic review. Addiction 107, 1 (2012), 39–50.

[119] LU, Y., WU, Y., LIU, J., LI, J., AND ZHANG, P. Understanding health care social media use from different stakeholder perspectives: a content analysis of an online health community. Journal of medical Internet research 19, 4 (2017), e109.

[120] LUNDBERG, S. M., AND LEE, S.-I. A unified approach to interpreting model predictions. Advances in neural information processing systems 30 (2017).

[121] LWIN, M. O., LU, J., SHELDENKAR, A., SCHULZ, P. J., SHIN, W., GUPTA, R., AND YANG, Y. Global sentiments surrounding the covid-19 pandemic on twitter: analysis of twitter trends. JMIR public health and surveillance 6, 2 (2020), e19447.

[122] MA, L., AND WANG, Y. Constructing a semantic graph with depression symptoms extraction from twitter. In 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) (2019), IEEE, pp. 1–5.

[123] MAHDIKHANI, M. Predicting the popularity of tweets by analyzing public opinion and emotions in different stages of covid-19 pandemic. International Journal of Information Management Data Insights 2, 1 (2022), 100053.

[124] MALAWSKI, M. A note on equal treatment and symmetry of values. In Transactions on Computational Collective Intelligence XXXV (2020), Springer, pp. 76–84.

[125] MARINAI, S., AND DENGEL, A. Document Analysis Systems VI: 6th International Workshop, DAS 2004, Florence, Italy, September 8-10, 2004, Proceedings, vol. 3163. Springer, 2004.

[126] MARKOULIDAKIS, I., KOPSIAFTIS, G., RALLIS, I., AND GEORGOULAS, I. Multi-class confusion matrix reduction method and its application on net promoter score classification problem. In The 14th pervasive technologies related to assistive environments conference (2021), pp. 412–419.

[127] MASHHADI, A., WINDER, S. G., LIA, E. H., AND WOOD, S. A. No walk in the park: The viability and fairness of social media analysis for parks and recreational policy making. In Proceedings of the International AAAI Conference on Web and Social Media (2021), vol. 15, pp. 409–420.

[128] MCCLELLAN, C., ALI, M. M., MUTTER, R., KROUTIL, L., AND LANDWEHR, J. Using social media to monitor mental health discussions- evidence from twitter. Journal of the American Medical Informatics Association 24, 3 (2017), 496–502.

[129] MEESALA, S. R., AND SUBRAMANIAN, S. Feature based opinion analysis on social media tweets with association rule mining and multi-objective evolutionary algorithms. Concurrency and Computation: Practice and Experience 34, 3 (2022), e6586.

[130] MEHRABI, N., GUPTA, U., MORSTATTER, F., STEEG, G. V., AND GALSTYAN, A. Attributing fair decisions with attention interventions. arXiv preprint arXiv:2109.03952 (2021).

[131] MEHRABI, N., MORSTATTER, F., SAXENA, N., LERMAN, K., AND GALSTYAN, A. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) 54, 6 (2021), 1–35.

[132] MENDES, R., AND VILELA, J. P. Privacy-preserving data mining: methods, metrics, and applications. IEEE Access 5 (2017), 10562–10582.

[133] MENDHE, C. H., HENDERSON, N., SRIVASTAVA, G., AND MAGO, V. A scalable platform to collect, store, visualize, and analyze big data in real time. IEEE Transactions on Computational Social Systems 8, 1 (2020), 260–269.

[134] MENGISTIE, T. T., ET AL. Covid-19 outbreak data analysis and prediction modeling using data mining technique. International Journal of Computer (IJC) 38, 1 (2020), 37–60.

[135] MILIOU, I., PAVLOPOULOS, J., AND PAPAPETROU, P. Sentiment nowcasting during the covid-19 pandemic. In Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings 24 (2021), Springer, pp. 218–228.

[136] MITCHELL, A., SHEARER, E., AND STOCKING, G. News on twitter: Consumed by most users and trusted by many. Pew Research Center (2021).

[137] MITTELSTADT, B. Principles alone cannot guarantee ethical ai. Nature machine intelligence 1, 11 (2019), 501–507.

[138] MOORES, B., AND MAGO, V. A survey on automated sarcasm detection on twitter. arXiv preprint arXiv:2202.02516 (2022).

[139] MURDOCH, W. J., SINGH, C., KUMBIER, K., ABBASI-ASL, R., AND YU, B. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences 116, 44 (2019), 22071–22080.

[140] NABIL, M., ALY, M., AND ATIYA, A. Astd: Arabic sentiment tweets dataset. In Proceedings of the 2015 conference on empirical methods in natural language processing (2015), pp. 2515–2519.

[141] NARASIMHAN, H., COTTER, A., GUPTA, M., AND WANG, S. Pairwise fairness for ranking and regression. In Proceedings of the AAAI Conference on Artificial Intelligence (2020), vol. 34, pp. 5248–5255.

[142] NAWAZ, M. S., BILAL, M., LALI, M. I., UL MUSTAFA, R., ASLAM, W., AND JAJJA, S. Effectiveness of social media data in healthcare communication. Journal of Medical Imaging and Health Informatics 7, 6 (2017), 1365–1371.

[143] NEBEKER, C., PARRISH, E. M., AND GRAHAM, S. The ai artificial intelligence (ai)-powered digital health digital health sector: Ethical and regulatory considerations when developing digital mental health digital mental health mental health tools

for the older adult older adults demographic. In Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues. Springer, 2022, pp. 159–176.

[144] NEWMAN, D., LAU, J. H., GRIESER, K., AND BALDWIN, T. Automatic evaluation of topic coherence. In Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics (2010), pp. 100–108.

[145] NGUYEN, A. T., RAFF, E., NICHOLAS, C., AND HOLT, J. Leveraging uncertainty for improved static malware detection under extreme false positive constraints. arXiv preprint arXiv:2108.04081 (2021).

[146] NOGUEIRA, A. R., PUGNANA, A., RUGGIERI, S., PEDRESCHI, D., AND GAMA, J. Methods and tools for causal discovery and causal inference. Wiley interdisciplinary reviews: data mining and knowledge discovery 12, 2 (2022), e1449.

[147] NUSHI, B., KAMAR, E., AND HORVITZ, E. Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (2018), vol. 6, pp. 126–135.

[148] ORGANIZATION, W. H., ET AL. Medicines transparency alliance (meta): pathways to transparency, accountability an access: cross-case analysis and review of phase ii.

[149] OZGA, J. The politics of accountability. Journal of Educational Change 21, 1 (2020), 19–35.

[150] PAREDES, J. N., TEZE, J. C. L., MARTINEZ, M. V., AND SIMARI, G. I. The heic application framework for implementing xai-based socio-technical systems. Online Social Networks and Media 32 (2022), 100239.

[151] PARK, H. W., PARK, S., AND CHONG, M. Conversations and medical news frames on twitter: Infodemiological study on covid-19 in south korea. Journal of medical internet research 22, 5 (2020), e18897.

[152] PARK, Y., HU, J., SINGH, M., SYLLA, I., DANKWA-MULLAN, I., KOSKI, E., AND DAS, A. K. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. JAMA network open 4, 4 (2021), e213909–e213909.

[153] PARKER, C. Meta-regulation: legal accountability for corporate social responsibility.

[154] PASTALTZIDIS, I., DIMITRIOU, N., QUEZADA-TAVAREZ, K., AIDINLIS, S., MARQUENIE, T., GURZAWSKA, A., AND TZOVARAS, D. Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. In 2022 ACM Conference on Fairness, Accountability, and Transparency (2022), pp. 2302–2314.

[155] PEISENIEKS, J., AND SKADIŅŠ, R. Uses of machine translation in the sentiment analysis of tweets. In Human Language Technologies–The Baltic Perspective. IOS Press, 2014, pp. 126–131.

[156] PENCINA, M. J., D'AGOSTINO SR, R. B., AND DEMLER, O. V. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. Statistics in medicine 31, 2 (2012), 101–113.

[157] PERSHAD, Y., HANGGE, P. T., ALBADAWI, H., AND OKLU, R. Social medicine: Twitter in healthcare. Journal of clinical medicine 7, 6 (2018), 121.

[158] PIRRAGLIA, P. A., AND KRAVITZ, R. L. Social media: new opportunities, new ethical concerns. Journal of general internal medicine 28 (2013), 165–166.

[159] PODDAR, S., MONDAL, M., MISRA, J., GANGULY, N., AND GHOSH, S. Winds of change: Impact of covid-19 on vaccine-related opinions of twitter users. In Proceedings of the International AAAI Conference on Web and Social Media (2022), vol. 16, pp. 782–793.

[160] RAIHAN, M., ISLAM, M. T., GHOSH, P., HASSAN, M. M., ANGON, J. H., AND KABIRAJ, S. Human behavior analysis using association rule mining techniques. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (2020), IEEE, pp. 1–5.

[161] RAJI, I. D., SMART, A., WHITE, R. N., MITCHELL, M., GEBRU, T., HUTCHINSON, B., SMITH-LOUD, J., THERON, D., AND BARNES, P. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 conference on fairness, accountability, and transparency (2020), pp. 33–44.

[162] RAZIS, G., AND ANAGNOSTOPOULOS, I. Influencetracker: Rating the impact of a twitter account. In Artificial Intelligence Applications and Innovations: AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, September 19-21, 2014. Proceedings 10 (2014), Springer, pp. 184–195.

[163] REICH, M. R. The core roles of transparency and accountability in the governance of global health public–private partnerships. Health Systems & Reform 4, 3 (2018), 239–248.

[164] RÖDER, M., BOTH, A., AND HINNEBURG, A. Exploring the space of topic coherence measures. In Proceedings of the eighth ACM international conference on Web search and data mining (2015), pp. 399–408.

[165] ROSENBERG, H., SYED, S., AND REZAIE, S. The twitter pandemic: The critical role of twitter in the dissemination of medical information and misinformation during the covid-19 pandemic. Canadian journal of emergency medicine 22, 4 (2020), 418–421.

[166] ROSENFELD, A., AND RICHARDSON, A. Explainability in human–agent systems. Autonomous Agents and Multi-Agent Systems 33 (2019), 673–705.

[167] RUFAI, S. R., AND BUNCE, C. World leaders' usage of twitter in response to the covid-19 pandemic: a content analysis. Journal of public health 42, 3 (2020), 510–516.

[168] RUSTAM, F., KHALID, M., ASLAM, W., RUPAPARA, V., MEHMOOD, A., AND CHOI, G. S. A performance comparison of supervised machine learning models for covid-19 tweets sentiment analysis. Plos one 16, 2 (2021), e0245909.

[169] SAHA, D., SCHUMANN, C., MCELFRESH, D., DICKERSON, J., MAZUREK, M., AND TSCHANTZ, M. Measuring non-expert comprehension of machine learning fairness metrics. In International Conference on Machine Learning (2020), PMLR, pp. 8377–8387.

[170] SALEIRO, P., KUESTER, B., HINKSON, L., LONDON, J., STEVENS, A., ANISFELD, A., RODOLFA, K. T., AND GHANI, R. Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577 (2018).

[171] SANFEY, A. G. Social decision-making: insights from game theory and neuroscience. Science 318, 5850 (2007), 598–602.

[172] SAXENA, N. A., HUANG, K., DEFILIPPIS, E., RADANOVIC, G., PARKES, D. C., AND LIU, Y. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (2019), pp. 99–106.

[173] SHAH, N., SRIVASTAVA, G., SAVAGE, D. W., AND MAGO, V. Assessing canadians health activity and nutritional habits through social media. Frontiers in public health 7 (2020), 400.

[174] SHARMA, S. Data privacy and GDPR handbook. John Wiley & Sons, 2019.

[175] SHNEIDERMAN, B. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems. ACM Transactions on Interactive Intelligent Systems (TiiS) 10, 4 (2020), 1–31.

[176] SINGH, R., AND SINGH, R. Applications of sentiment analysis and machine learning techniques in disease outbreak prediction–a review. Materials Today: Proceedings (2021).

[177] SINGHAL, A., BAXI, M. K., MAGO, V., ET AL. Synergy between public and private health care organizations during covid-19 on twitter: Sentiment and engagement analysis using forecasting models. JMIR Medical Informatics 10, 8 (2022), e37829.

[178] SINGHAL, A., TANVEER, H., AND MAGO, V. Towards fate in ai for social media and healthcare: A systematic review. arXiv preprint arXiv:2306.05372 (2023).

[179] SLACK, D., FRIEDLER, S. A., SCHEIDEGGER, C., AND ROY, C. D. Assessing the local interpretability of machine learning models. arXiv preprint arXiv:1902.03501 (2019).

[180] SLAVIK, C. E., BUTTLE, C., STURROCK, S. L., DARLINGTON, J. C., AND YIANNAKOULIAS, N. Examining tweet content and engagement of canadian public health agencies and decision makers during covid-19: mixed methods analysis. Journal of Medical Internet Research 23, 3 (2021), e24883.

[181] SOKOL, K., AND FLACH, P. A. Counterfactual explanations of machine learning predictions: Opportunities and challenges for ai safety. SafeAI@ AAAI (2019).

[182] SOMEH, I., DAVERN, M., BREIDBACH, C. F., AND SHANKS, G. Ethical issues in big data analytics: A stakeholder perspective. Communications of the Association for Information Systems 44, 1 (2019), 34.

[183] SON, J., LEE, J., OH, O., LEE, H. K., AND WOO, J. Using a heuristic-systematic model to assess the twitter user profile's impact on disaster tweet credibility. International Journal of Information Management 54 (2020), 102176.

[184] SØRENSEN, K. Health literacy: A key attribute for urban settings. Optimizing Health Literacy for Improved Clinical Practices (2018), 1–16.

[185] STANBERRY, B. Legal and ethical aspects of telemedicine. Journal of telemedicine and telecare 12, 4 (2006), 166–175.

[186] STELLEFSON, M., PAIGE, S. R., CHANEY, B. H., AND CHANEY, J. D. Evolving role of social media in health promotion: updated responsibilities for health education specialists. International journal of environmental research and public health 17, 4 (2020), 1153.

[187] STEPIN, I., ALONSO, J. M., CATALA, A., AND PEREIRA-FARIÑA, M. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access 9 (2021), 11974–12001.

[188] SWAN, M. Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking. International journal of environmental research and public health 6, 2 (2009), 492–525.

[189] TAN, P.-N., STEINBACH, M., AND KUMAR, V. Introduction to data mining. Pearson Education India, 2016.

[190] TANG, L., LIU, W., THOMAS, B., TRAN, H. T. N., ZOU, W., ZHANG, X., AND ZHI, D. Texas public agencies' tweets and public engagement during the covid-19 pandemic: Natural language processing approach. JMIR public health and surveillance 7, 4 (2021), e26720.

[191] TAO, G., SUN, W., HAN, T., FANG, C., AND ZHANG, X. Ruler: discriminative and iterative adversarial training for deep neural network fairness. In Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (2022), pp. 1173–1184.

[192] TASSONE, J., YAN, P., SIMPSON, M., MENDHE, C., MAGO, V., AND CHOUDHURY, S. Utilizing deep learning and graph mining to identify drug use on twitter data. BMC Medical Informatics and Decision Making 20, 11 (2020), 1–15.

[193] TOMMASEL, A., DIAZ-PACE, A., RODRIGUEZ, J. M., AND GODOY, D. Forecasting mental health and emotions based on social media expressions during the covid-19 pandemic. Information Discovery and Delivery 49, 3 (2021), 259–268.

[194] TYRAWSKI, J., AND DEANDREA, D. C. Pharmaceutical companies and their drugs on social media: a content analysis of drug information on popular social media sites. Journal of medical Internet research 17, 6 (2015), e130.

[195] UMBRELLO, S., AND VAN DE POEL, I. Mapping value sensitive design onto ai for social good principles. AI and Ethics 1, 3 (2021), 283–296.

[196] UNERMAN, J. Stakeholder engagement and dialogue. In Sustainability accounting and accountability. Routledge, 2010, pp. 105–122.

[197] VALKO, M., AND HAUSKRECHT, M. Feature importance analysis for patient management decisions. Studies in health technology and informatics 160, Pt 2 (2010), 861.

[198] VENTOLA, C. L. Social media and health care professionals: benefits, risks, and best practices. Pharmacy and therapeutics 39, 7 (2014), 491.

[199] VERGEER, P., VAN SCHAIK, Y., AND SJERPS, M. Measuring calibration of likelihood-ratio systems: a comparison of four metrics, including a new metric devpav. Forensic Science International 321 (2021), 110722.

[200] VESNIC-ALUJEVIC, L., NASCIMENTO, S., AND POLVORA, A. Societal and ethical impacts of artificial intelligence: Critical notes on european policy frameworks. Telecommunications Policy 44, 6 (2020), 101961.

[201] WACHTER, S., MITTELSTADT, B., AND FLORIDI, L. Transparent, explainable, and accountable ai for robotics. Science robotics 2, 6 (2017), eaan6080.

[202] WANG, H., LI, Y., HUTCH, M., NAIDECH, A., LUO, Y., ET AL. Using tweets to understand how covid-19–related health beliefs are affected in the age of social media: Twitter data analysis study. Journal of medical Internet research 23, 2 (2021), e26302.

[203] WEISKOPF, N. G., HRIPCSAK, G., SWAMINATHAN, S., AND WENG, C. Defining and measuring completeness of electronic health records for secondary use. Journal of biomedical informatics 46, 5 (2013), 830–836.

[204] WEISS, D. J., NELSON, A., GIBSON, H., TEMPERLEY, W., PEEDELL, S., LIEBER, A., HANCHER, M., POYART, E., BELCHIOR, S., FULLMAN, N., ET AL. A global map of travel time to cities to assess inequalities in accessibility in 2015. Nature 553, 7688 (2018), 333–336.

[205] WENG, T.-W., ZHANG, H., CHEN, P.-Y., YI, J., SU, D., GAO, Y., HSIEH, C.-J., AND DANIEL, L. Evaluating the robustness of neural networks: An extreme value theory approach. arXiv preprint arXiv:1801.10578 (2018).

[206] WIBOWO, W., SARI, N. P., WILANTARI, R. N., AND ABDUL-RAHMAN, S. Association rule mining method for the identification of internet use. In Journal of Physics: Conference Series (2021), vol. 1874, IOP Publishing, p. 012009.

[207] WIERINGA, M. What to account for when accounting for algorithms. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (2020), pp. 1–18.

[208] WRIGHT, D. A framework for the ethical impact assessment of information technology. Ethics and information technology 13 (2011), 199–226.

[209] XIONG, Z., CUI, Y., LIU, Z., ZHAO, Y., HU, M., AND HU, J. Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. Computational Materials Science 171 (2020), 109203.

[210] XU, J., XIAO, Y., WANG, W. H., NING, Y., SHENKMAN, E. A., BIAN, J., AND WANG, F. Algorithmic fairness in computational medicine. EBioMedicine 84 (2022), 104250.

[211] XUE, J., CHEN, J., HU, R., CHEN, C., ZHENG, C., SU, Y., AND ZHU, T. Twitter discussions and emotions about the covid-19 pandemic: Machine learning approach. Journal of medical Internet research 22, 11 (2020), e20550.

[212] YAO, H., CHEN, Y., YE, Q., JIN, X., AND REN, X. Refining language models with compositional explanations. Advances in Neural Information Processing Systems 34 (2021), 8954–8967.

[213] YAO, S., AND HUANG, B. Beyond parity: Fairness objectives for collaborative filtering. Advances in neural information processing systems 30 (2017).

[214] ZAFAR, M. B., VALERA, I., GOMEZ-RODRIGUEZ, M., AND GUMMADI, K. P. Fairness constraints: A flexible approach for fair classification. The Journal of Machine Learning Research 20, 1 (2019), 2737–2778.

[215] ZAFAR, M. R., AND KHAN, N. Deterministic local interpretable model-agnostic explanations for stable explainability. Machine Learning and Knowledge Extraction 3, 3 (2021), 525–541.

[216] ZAKI, M. M., JENA, A. B., AND CHANDRA, A. Supporting value-based health care-aligning financial and legal accountability. The New England journal of medicine 385, 11 (2021), 965–967.

[217] ZHAI, C., COHEN, W. W., AND LAFFERTY, J. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In Acm sigir forum (2015), vol. 49, ACM New York, NY, USA, pp. 2–9.

[218] ZHANG, Y., AND ZHOU, L. Fairness assessment for artificial intelligence in financial industry. arXiv preprint arXiv:1912.07211 (2019).

[219] ZHAO, X., LOVREGLIO, R., AND NILSSON, D. Modelling and interpreting pre-evacuation decision-making using machine learning. Automation in Construction 113 (2020), 103140.

[220] ZHOU, J., LIU, F., AND ZHOU, H. Understanding health food messages on twitter for health literacy promotion. Perspectives in public health 138, 3 (2018), 173–179.

[221] ZHOU, L., ZHANG, D., YANG, C. C., AND WANG, Y. Harnessing social media for health information management. Electronic commerce research and applications 27 (2018), 139–151.

Chapter 6

Appendix

6.1 Topics and User Engagement

Table 6.1: Model parameters for topic clustering with TF-IDF document embeddings.

Clustering Algorithm	Epochs	Chunk Size	Workers (Number of CPU cores)	Evaluation Period (seconds)	(A-priori belief on document - topic distribution)	(A-priori belief on topic - word distribution)	(Gradient descent step size)	Minimum normalizing probability
LDA	50	1000	NA	10	0.01	0.9	NA	NA
Parallel LDA	50	1000	7	10	0.01	0.9	NA	NA
LSI	NA	1000	NA	NA	NA	NA	NA	NA
NMF	50	1000	NA	10	NA	NA	1	0
HDP	NA	1000	NA	NA	0.01	NA	1	NA

Table 6.2: Sample of topic keywords generated using HDP and NMF.

HDP	NMF
['commonwealth','speedy','multi-vitamin','vaccine', 'weather-wise','unopen','salmon','breadth','land', '#skincancerawarenessmonth']	['vaccine','disease','protect people','prevent death', 'cancer','research']
['prop','goldstein','mihcha','kezspm','age','open','mohmv', 'thisisdiabetic','onco']	['health for all','healthcare','community health', 'vaccines work']

Table 6.3: List of topics obtained using NMF model. Italicized topic keywords are repeated in both timeframes, before COVID-19 and during COVID-19.

Time Phase	Topic	Topic Keywords
Before COVID-19	*Health Research*	*['cancer','research','vaccine','advancements','find','medical research', 'national institute for health','nih research','icon pra','qualitative health research', 'mental health research','mhsrs','public health research','integrative medicine research', 'medical trials']*
	Community Healthcare	*['community health','care','community health services','health center', 'family health centers','community plan','community clinic','family healthcare', 'qualified health centers','health services']*
	Chronic Diseases	*['angina','arthritis','asthma','bipolar disorder','cancer','hypertension','stroke', 'COPD','diabetes','heart attack','sleep apnea','disease','chronic','lupus', 'multiplesclerosis','lung cancer','ovarian cancer','heart failure','kidney disease', 'breast cancer','prostate cancer','spinal disorder','hiv','hemophilia','pneumonia', 'malaria','aids','tb','tuberculosis']*
	Medical Trials	*['clinical trials','medical trials','paid trials','medical research studies', 'cancer clinical trials','hydoxychloroquine studies','randomized clinical trial', 'applied clinical trials','oncology clinical trials','celerion clinical trials', 'alzheimer clinical trials','registered clinical trials','depression clinical trials', 'weight loss clinical trials','artificial kidney human trials']*
	Customer Experience	['connected customer','customer experience','customer journey','user journey', 'user happiness','client satisfaction','seamless experience','measuring customer experience']
During COVID-19	COVID-19	['covid 19', 'virus', 'coronavirus', 'covid 19 cases', 'covid 19 deaths','covid 19 passport', 'covid 19 insurance', 'quarantine', 'pandemic', 'outbreak', 'social distancing', 'self isolation', 'cases', 'deaths', 'infections', 'fatality rate', 'mortality', 'masks', 'hygiene', 'state of emergency', 'surveillance', 'infectivity', 'communicable disease', 'community spread','containment','epidemic', 'herd immunity','ppe','personal protective equipment','respirator','SPO2', 'severe acute respiratory syndrome','contact tracing','hydroxychloroquine','risk']
	Vaccination	['covaxin','vaccine','covid vaccine','mrna vaccine','vaccine finder','herd immunity','booster shot', 'vaccine appointment','mandatory vaccine','vaccination card','vaccination passport', 'vaccination rates','inoculation','covishield']
	Mental Health	['anxiety','bipolar disorder','depression','panic','ptsd','schizophrenia','sucidal ideation', 'suicide','alzheimers','parkinson','mental illness','mental health day','mental health counselor', 'mental health services','mental disorder','clinical psychologist','behavioral health', 'mental health awareness','mental health therapist','mhfa','mental disability','psychologist', 'family therapists','licensed clinical social worker','strong minds','mental health stigma', 'mental health resources']
	Nutrition and Well-being	['healthy living','community','support','helping','awareness','development','innovation', 'well being','nutrition','diet','healthy diet','skin fuel','eat well be healthy', 'understanding nutrition and well being',' good sleep','nutritious foods']
	Community Healthcare	*['community health','care','community health services','health center','family health centers', 'community plan','community clinic','family healthcare','qualified health centers','health services']*
	Health Research	*['cancer','research','vaccine','advancements','find','medical research','national institute for health', 'nih research','icon pra','qualitative health research','mental health research','mhsrs', 'public health research','integrative medicine research','medical trials']*
	Chronic Diseases	*['angina','arthritis','asthma','bipolar disorder','cancer','hypertension','stroke','COPD','diabetes', 'heart attack','sleep apnea','disease','chronic','lupus','multiplesclerosis','lung cancer', 'ovarian cancer','heart failure','kidney disease','breast cancer','prostate cancer','spinal disorder',' hiv','hemophilia','pneumonia','malaria','aids','tb','tuberculosis']*
	Medical Trials	*['clinical trials','medical trials','paid trials','medical research studies','cancer clinical trials', 'hydoxychloroquine studies','randomized clinical trial','applied clinical trials', 'oncology clinical trials','celerion clinical trials','alzheimer clinical trials','registered clinical trials', 'depression clinical trials','weight loss clinical trials','artificial kidney human trials']*

Table 6.4: Selected tweets having high user engagement.

Organization	Tweet ID	Created at	Tweet	Average user engagement/ Average user engagement with impact
Pfizer	1325767629890592771	2020-11-09 11:50:09+00:00	UPDATE: We are proud to announce, along with @BioNTech_Group, that our mRNA-based #vaccine candidate has, at an interim analysis, demonstrated initial evidence of efficacy against #COVID19 in participants without prior evidence of SARS-CoV-2 infection.	13,901.75 / 319.74
Pfizer	1389203084879011840	2021-05-03 13:00:00+00:00	Today we have announced we are mobilizing the largest humanitarian relief effort in our company's history to help the people of India fight the vicious second wave of coronavirus that is currently ravaging the nation. https://t.co/klVnkAjkcw	3,132.25 / 72.04
CDC	1392911350058323973	2021-05-13 18:35:19+00:00	UPDATE: If you are fully vaccinated against #COVID19, you can resume activities without wearing a mask or staying 6 feet apart, except where required by federal, state, local, tribal or territorial laws, incl. local business and workplace guidance. More: https://t.co/FJMon7WlFO	28,997.50 / 20,124.26
WHO	[HTML]FFFFFF1313841832598687749	[HTML]FFFFFF2020-10-07 14:01:17+00:00	[HTML]FFFFFF We are thrilled to have @SuperM joining our Big Event for Mental Health on #WorldMentalHealthDay! Stay tuned for #SuperMxWHO! This Saturday 10.10.2020 10h00 EST 16h00 CEST 23h00 KST More information: https://t.co/seFE6mb3O7 #SuperM https://t.co/zfYaJfr6nn	17,398.00 / 17,398.00

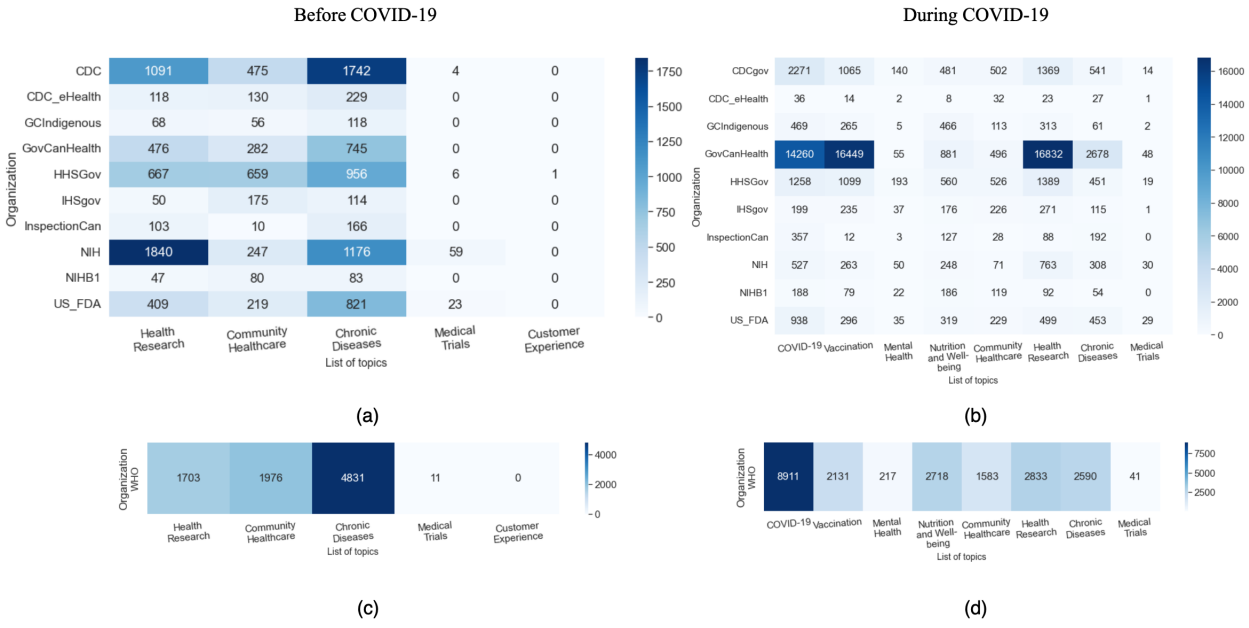


Figure 6.1: Scaled heatmaps showing topic distribution for Public Health Agencies and WHO before COVID-19 and during COVID-19.

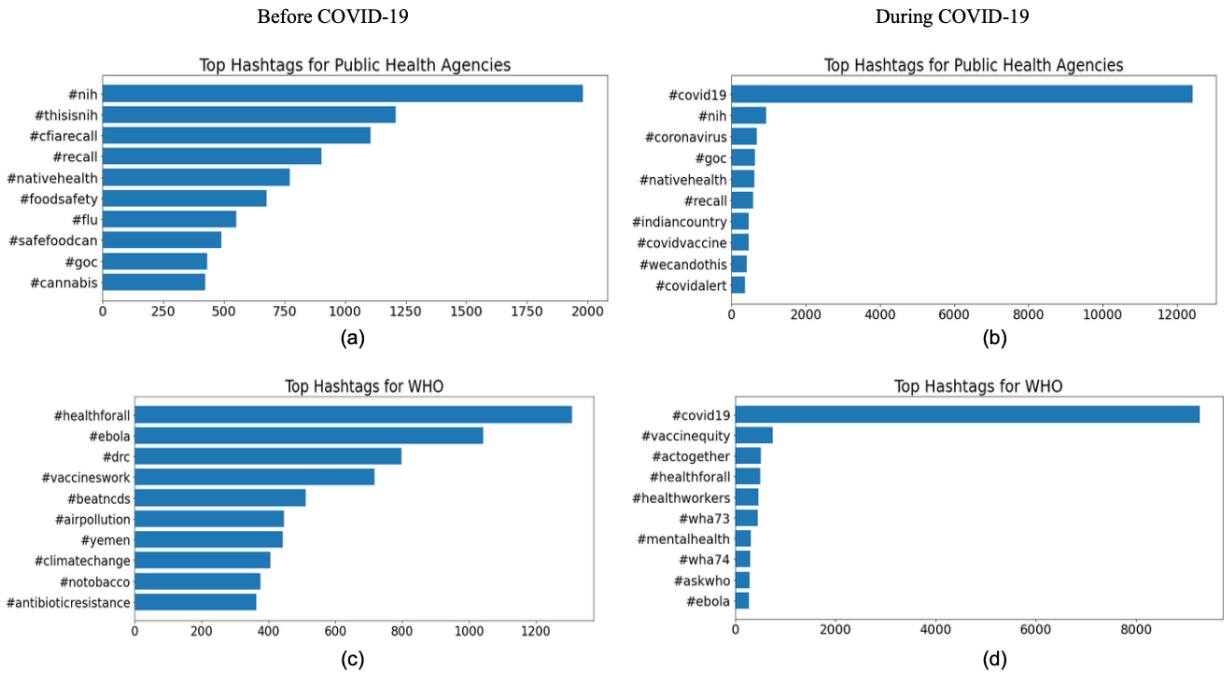


Figure 6.2: Top hashtags for different organizations before COVID-19 and during COVID-19 for Public Health Agencies and WHO.

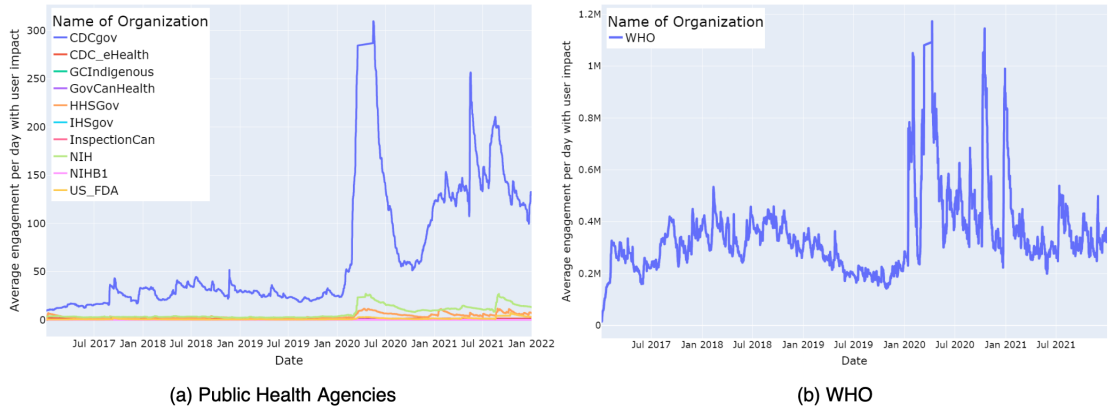


Figure 6.3: User Engagement on Twitter accounts of Public Health Agencies and WHO from January 01, 2017 to December 31, 2021.

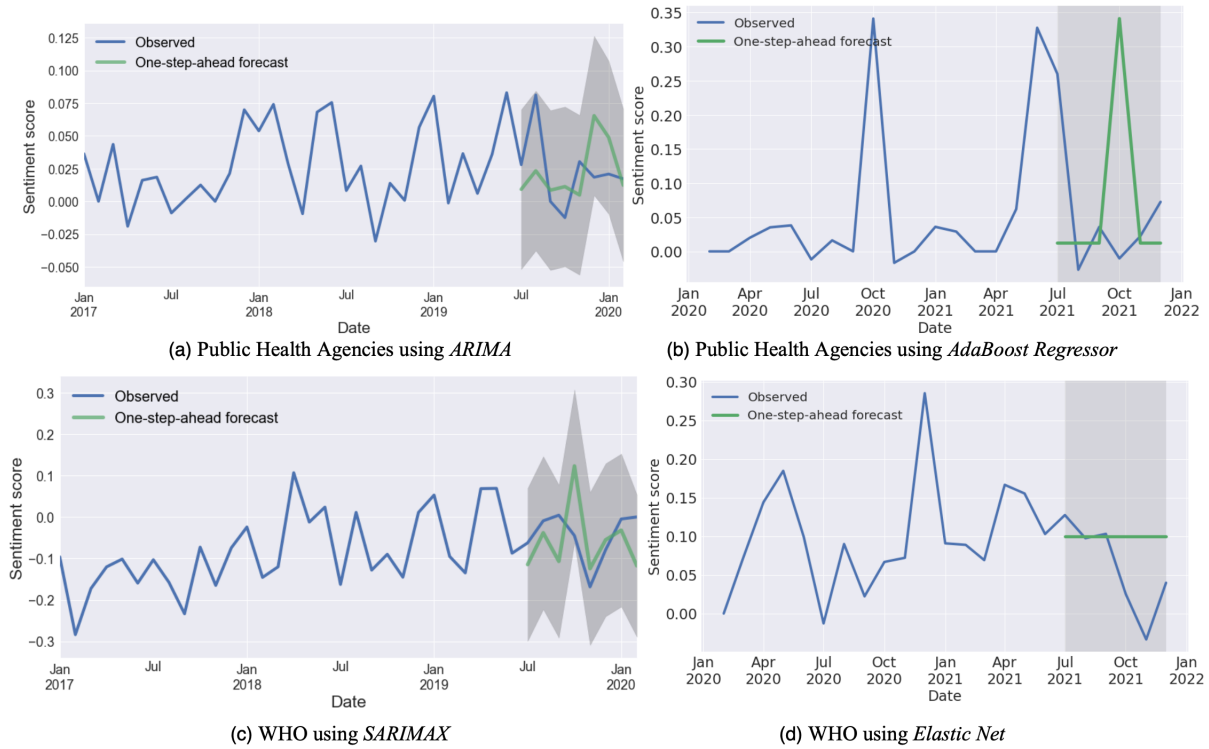


Figure 6.4: One-step ahead forecast for Public Health Agencies and WHO before COVID-19 and during COVID-19 using the best performing models from Table S4.

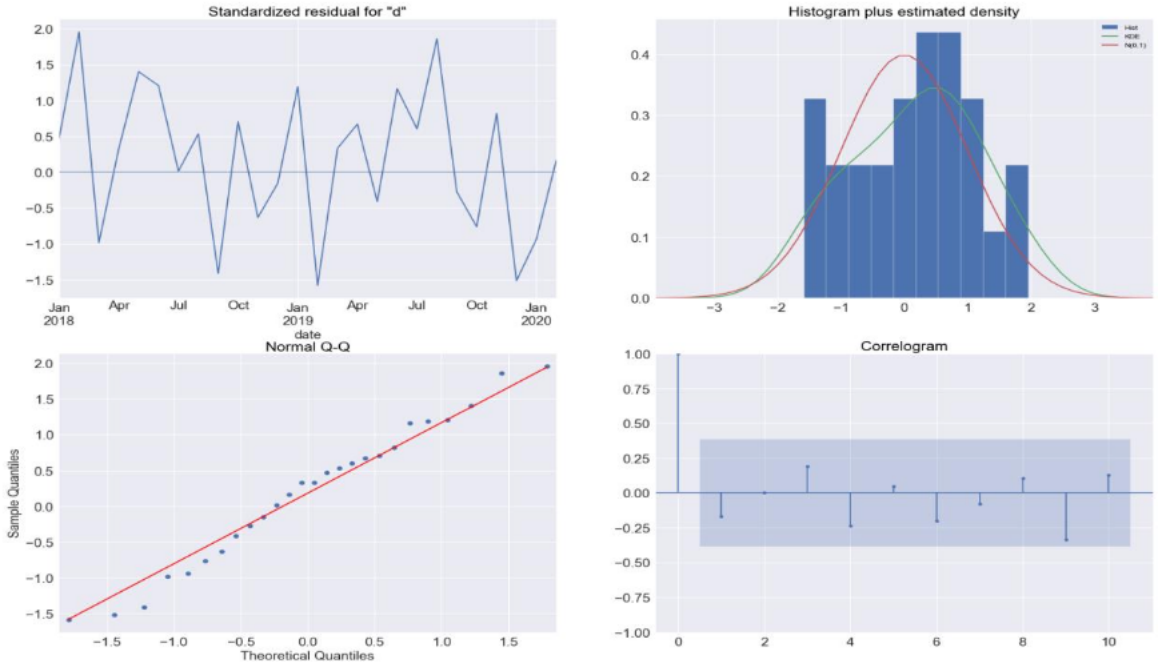


Figure 6.5: plot_diagnostics for Public Health Agencies before COVID-19 using ARIMA.

PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Page 1
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	Page 1
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Page 2
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Page 2
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Page 2
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Page 2
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Page 3
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Page 3
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Page 3
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	NA
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	NA
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	NA
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	NA
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Page 3
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	Page 3
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Pages 6, 9, 11, 13
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	NA
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	NA
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	NA
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	Page 15
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	Page 14
RESULTS			

PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Page 4
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	NA
Study characteristics	17	Cite each included study and present its characteristics.	Pages 1-21
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	NA
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	NA
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	NA
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	NA
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	NA
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	NA
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	NA
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	Sections 2-5
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Pages 14-15
	23b	Discuss any limitations of the evidence included in the review.	Page 15
	23c	Discuss any limitations of the review processes used.	NA
	23d	Discuss implications of the results for practice, policy, and future research.	Pages 15-16
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	NA
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	NA
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	NA
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	Acknowledgement
Competing interests	26	Declare any competing interests of review authors.	Page 16
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Page 2

From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71