UNIVERSITY OF CALIFORNIA,
IRVINE

Computational Analysis of Health Text

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Informatics

by

Lu He

Dissertation Committee:
Professor Kai Zheng, FACMI, Chair
Associate Professor Yunan Chen
Assistant Professor Daniel Epstein
Clinical Assistant Professor Helen Ma

2023

DEDICATION

To my grandparents: Xiufen Chu and Jiemin Wang

TABLE OF CONTENTS

Page

LIST OF FIGURES **vi**

LIST OF TABLES **vii**

ACKNOWLEDGMENTS **viii**

VITA **x**

ABSTRACT OF THE DISSERTATION **xii**

1 Introduction **1**
- 1.1 Health text data: an overview . . . 3
  - 1.1.1 The voices of patients and the public: health-related social media data 3
  - 1.1.2 Rich but dense: clinical texts . . . 8
- 1.2 Contributions . . . 12
- 1.3 Dissertation overview . . . 12

2 Study 1: Evaluating and improving the use of computational analysis of health-related social media data **14**
- 2.1 Study 1A: Developing a standardized protocol for computational sentiment analysis research using health-related social media data . . . 14
  - 2.1.1 Study summary . . . 14
  - 2.1.2 Introduction . . . 15
  - 2.1.3 Objective . . . 18
  - 2.1.4 Material and methods . . . 18
  - 2.1.5 Results . . . 20
  - 2.1.6 Discussion . . . 30
  - 2.1.7 Conclusion . . . 33
- 2.2 Study 1B: Empirical evaluation of computational sentiment analysis tools on health-related social media data . . . 34
  - 2.2.1 Study summary . . . 34
  - 2.2.2 Introduction . . . 35
  - 2.2.3 Materials and methods . . . 37
  - 2.2.4 Results . . . 41
  - 2.2.5 Discussion . . . 50

2.2.6 Conclusion . . . 53

3 Study 2: Developing computer-assisted qualitative analysis to understand public and patient concerns toward health-related issues 55

3.1 Study 2A: What do patients care about? Mining fine-grained patient concerns from online physician reviews through computer-assisted multi-level qualitative analysis . . . 55
3.1.1 Study summary . . . 56
3.1.2 Introduction . . . 56
3.1.3 Material and methods . . . 58
3.1.4 Results . . . 62
3.1.5 Discussion . . . 70
3.1.6 Conclusion . . . 72
3.2 Study 2B: Why do people oppose mask wearing? A comprehensive analysis of U.S. tweets during the COVID-19 pandemic . . . 73
3.2.1 Study summary . . . 73
3.2.2 Introduction . . . 74
3.2.3 Material and methods . . . 76
3.2.4 Results . . . 83
3.2.5 Discussion . . . 89
3.2.6 Conclusion . . . 94

4 Study 3: Developing and validating a Natural Language Processing (NLP) pipeline for clinical information extraction from notes of veterans with lymphoid malignancies 96

4.1 Study summary . . . 96
4.2 Introduction . . . 97
4.3 Material and methods . . . 98
4.4 Results . . . 102
4.4.1 Descriptive summary . . . 102
4.4.2 NLP performance . . . 102
4.4.3 Error analysis . . . 103
4.5 Discussion . . . 104
4.6 Conclusion . . . 106

5 Discussion and future work 107

5.1 Discussion . . . 107
5.1.1 Summary of dissertation research . . . 107
5.1.2 Implications for computational analysis of health text in the LLM era 111
5.2 Future work . . . 113
5.2.1 Integrating large language models into health text analysis . . . 113
5.2.2 Addressing biases and disparities in computational analysis of health text . . . 114
5.2.3 Embracing multimodal health data . . . 115
5.2.4 Implementing data-driven systems in real-world settings . . . 116

6 Concluding Remarks **117**

Bibliography **118**

Appendix A **137**

LIST OF FIGURES

Page

2.1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Flow Diagram. 21
2.2 PATH 29
2.3 Weighted F1 scores. 42
2.4 Overestimation and underestimation of the tools on the datasets. 45
2.5 Fleiss' Kappa scores on the seven datasets. 46
2.6 Lexicon composition. 48
2.7 Overlapping among sentiment lexicons. 48

3.1 Method pipeline. 62
3.2 Topic distribution in different ratings. 65
3.3 Aspect distribution in different topics and ratings. 70
3.4 Method flowchart. RQ: research question. 80
3.5 Temporal change of tweet volume and attitude. CDC: Centers for Disease Control and Prevention. 84

4.1 Method flowchart. 101

LIST OF TABLES

Page

2.1 Search query 19
2.2 Distinctive research design choices 26
2.3 Methodological and reporting inconsistencies among the existing studies (N = 89) 28
2.4 The seven sentiment lexicons evaluated. 38
2.5 The four sentiment analyzers evaluated. 38
2.6 Description of the validation social media datasets. 40
2.7 Average weighted-F1 score by tool types and data types. 46

3.1 Coarse-grained topics. 64
3.2 Aspects under Relationship. 66
3.3 Aspects under Clinical. 68
3.4 Aspects under Management. 69
3.5 Search keywords 77
3.6 Evolution of frequently discussed topics over time 85
3.7 Major categories of concerns or justifications for opposing mask wearing (examples paraphrased to protect confidentiality) 87
3.8 Comparison of use of external information among pro- vs anti-mask tweets . 89

4.1 NLP performance on the test set with 131 notes. 102

ACKNOWLEDGMENTS

I thank my advisor, Dr. Kai Zheng, for supporting me in every possible way and always providing sharp but constructive feedback. He is the inspiration for me to stay in academia and become an advisor like him in the future.

I thank Dr. Yunan Chen for introducing me to the world of Human-Computer Interaction and always pushing me to think deeper and more critically. I thank Dr. Helen Ma for tirelessly teaching me mini-lectures on lymphoid malignancies and clinicians' work. I thank Dr. Daniel Epstein for agreeing to serve on my comprehensive committee and dissertation committee and always providing constructive feedback.

One of the most enjoyable things I experienced during my PhD is working with my fantastic collaborators. I had the honor to collaborate with wonderful researchers: Dr. Yang Gong, Dr. Ashley Griffin, Changyang He, Dr. Sherwin Kuo, Dr. Jonathan Quang, Dr. Lixin Song, Dr. Ting Song, Dr. Yue Wang, Dr. Ping Yu.

I thank my friends in Irvine: Dr. Clara Caldeira, Dr. Yao Du, Dr. Elizabeth Eikey, Dr. Mayara Costa Figueiredo, Mengqi Gao, Katie Genuario, Tian Guan, Dr. Xinning Gui, Yawen Guo, Zhaoxian Hu, Yicong Huang, Dr. Mustafa Hussain, Yuxin Liu, Xi Lu, Dr. Joanne Ly, Julie Oh, Chenying Qin, Dr. Tera Reynolds, Lucas de Melo Silva, Nick Su, Brian Tran, Dr. Pei Wang, Dr. Tao Wang, Wenhao Wang, Yang Yue, Chaeyoon Yoo, Weijun Yuan, Dr. June Zhao, Dr. Yunxia Zhao, Rachael Zehrung, Xinxin Zheng, Yuxiang Zi.

I also thank the continued support from my college friends, Ning Jiang and Xiufen Xu. Even though we don't see each other often, our little chat group is the source of happiness for me and always made me laugh. I also thank my friend Bree Zhang for always being my emotional support and spending enjoyable moments together.

I thank all my family members, especially my parents Shujie Wang and Jianxin He, and my cousin, Xiaozhe Luo.

I thank my partner, Zhendong, for taking care of everything while I am busy working or worrying. Life is better with you and the food and coffee you make. I can't wait to see what's ahead for us and Lucifer.

Last, I dedicate this dissertation to my grandparents. I miss you everyday and I wish you could see this.

Portions of Chapter 2 were originally published in the Journal of American Medical Informatics Association [93], used with permission from Oxford University Press. The co-authors listed in this publication are Tingjue Yin, Zhaoxian Hu, Yunan Chen, David A. Hanauer, and Kai Zheng. Lu He, Tingjue Yin, and Zhaoxian Hu conducted the systematic review of the literature. Lu He and Kai Zheng designed the study and drafted the manuscript. Yunan Chen and David A. Hanauer assisted in designing the study and reviewed and revised the manuscript. The work was supported in part by the National Center for Research Resources

and the National Center for Advancing Translational Sciences of the National Institutes of Health through grant UL1TR001414. Kai Zheng directed and supervised research which forms the basis for the dissertation.

Portions of Chapter 2 were originally published in the Journal of Biomedical Informatics [94], used with permission from Elsevier. The co-authors listed in this publication are Tingjue Yin and Kai Zheng. Tingjue Yin assisted in data analysis and Kai Zheng contributed to study design and manuscript writing. Kai Zheng directed and supervised research which forms the basis for the dissertation. I thank Dr. Kang Zhao and Dr. Xiangyu Wang for kindly sharing their data to support the evaluation conducted in this study.

Portions of Chapter 3 were originally published in the 2020 Proceeding of the American Medical Informatics Association [92], used with permission from the American Medical Informatics Association. The co-authors listed in this publication are Changyang He, Yue Wang, Zhaoxian Hu, Yunan Chen, and Kai Zheng. Changyang He contributed to data analysis and manuscript writing, Yue Wang assisted in study design and manuscript editing, and Yunan Chen and Kai Zheng contributed to study design and manuscript editing. Kai Zheng directed and supervised research which forms the basis for the dissertation.

Portions of Chapter 3 were originally published in the Journal of American Medical Informatics Association [91], used with permission from Oxford University Press. The co-authors listed in this publication are Changyang He, Yunan Chen, Tera L. Reynolds, Yicong Huang, Qiushi Bai, Chen Li, and Kai Zheng. Lu He, Changyang He, and Yunan Chen designed the study. Lu He annotated data, designed coding scheme, and drafted the manuscript. Changyang He annotated data, built machine learning models, and drafted a significant portion of the manuscript. Tera L. Reynolds resolved disagreement during qualitative analysis, assisted in literature review, and revised the manuscript. Kai Zheng contributed to the study design and revised the manuscript. Yicong Huang, Qiushi Bai, and Chen Li collected the data and provided feedback to the manuscript. Kai Zheng directed and supervised research which forms the basis for the dissertation.

Portions of Chapter 4 are currently in preparation to be submitted to a peer-reviewed journal. The co-authors listed are: Matthew Moldenhauer, Helen Ma, and Kai Zheng. Matthew Moldenhauer contributed to data annotation, Helen Ma contributed to data acquisition, data annotation, study design, and manuscript editing, and Kai Zheng contributed to study design and manuscript editing. Kai Zheng supervised research which forms the basis for the dissertation.

VITA

Lu He

EDUCATION

Doctor of Philosophy in Informatics **2023**
University of California, Irvine *Irvine, CA*

Bachelor of Science in Computer Science **2017**
University of Minnesota, Twin Cities *Minneapolis, MN*

RESEARCH EXPERIENCE

Graduate Research Assistant **2017–2023**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **2017–2018**
University of California, Irvine *Irvine, CA*

REFEREED JOURNAL PUBLICATIONS

Griffin A, **He L**, Sunjaya A, King A, Khan Z, Douthit B, Nwadiugw M, Subbin V, Braunstein M, Nguyen V, Jaffe C, Schleyer T. Clinical, technical, and implementation characteristics of real-world health applications using FHIR. Journal of American Medical Informatics Association Open. 2022;5. DOI: 10.1093/jamiaopen/ooac077

He L, Yin T, Zheng K. They may not work! An evaluation of eleven sentiment analysis tools on seven social media datasets. Journal of Biomedical Informatics. 2022;132:104142. PMID: 35835437

He L, He C. Help me #DebunkThis: unpacking individual and community's collaborative work in information credibility assessment. Proceedings of the ACM on Human-Computer Interaction, CSCW. 2022;6. DOI: 10.1145/3555138

Su Z, **He L**, Jariwala SP, Zheng K, Chen Y. "What is your envisioned future?": towards human-AI enrichment in data work of asthma care. Proceedings of the ACM on Human-Computer Interaction, CSCW. 2022;6. DOI: 10.1145/3555157

He C, **He L**, Lu T, Li B. Beyond entertainment: unpacking Danmaku and comments' role of information sharing and sentiment expression in online crisis videos. Proceedings of the ACM on Human-Computer Interaction, CSCW. 2021;5. DOI: 10.1145/3479555

He L*, He C*, Reynolds TL, Bai Q, Huang Y, Li C, Zheng K, Chen Y. Why do people oppose mask wearing? A comprehensive analysis of US tweets during the COVID-19 pandemic. Journal of American Medical Informatics Association. 2021;28(7):1564–73. PMCID: PMC7989302 (* equal contribution) (**Editor's Choice and Featured Article**)

He L, Yin T, Hu Z, Chen Y, Hanauer DA, Zheng K. Developing a standardized protocol for computational sentiment analysis research using health-related social media data. Journal of American Medical Informatics Association. 2021;28(6):1125–34. PMCID: PMC8200276

Ma H, Smith C.E, **He L**, Narayanan S, Giaquinto R.A, Evans R, Hanson L, Yarosh S. Write for life: persisting in online health communities through expressive writing and social support. Proceedings of the ACM on Human-Computer Interaction, CSCW. 2017;1. DOI: 10.1145/3134708

REFEREED CONFERENCE PUBLICATIONS

He C, **He L**, Lu Z, Li B. Seeking love and companionship through streaming: unpacking livestreamer-moderated senior matchmaking in China. In: Proceedings of the 2023 ACM Conference on Human Factors in Computing Systems (CHI '23) DOI: 10.1145/3544548.3581195

Guo Y, Zhu J, Huang Y, **He L**, He C, Li C, Zheng K. Public opinions toward COVID-19 vaccine mandates: a machine learning-based analysis of U.S. tweets. American Medical Informatics Association Annual Symposium Proceedings. 2022;502–511. PMCID: PMC10148373 (**Student Paper Competition Second Place**)

He L*, Song T*, Jiang Y, Yu P, Song L, Gong Y. To improve supportive care for patients taking oral anticancer agents. In: Proceedings of the 2021 World Congress on Health and Biomedical Informatic (MEDINFO '21) 2021;290:547–51. PMID: 35673076 (* equal contribution)

He L, He C, Wang Y, Hu Z, Zheng K, Chen Y. What do patients care about? Mining fine-grained patient concerns from online physician reviews through computer-assisted multi-level qualitative analysis. American Medical Informatics Association Annual Symposium Proceedings. 2020;544–53. PMCID: PMC8075539 (**Student Paper Competition Finalist**)

He L, Zheng K. How do general-purpose sentiment analyzers perform when applied to health-related online social media data? In: Proceedings of the 2019 World Congress on Health and Biomedical Informatics (MEDINFO '19). 2019;1208–12. PMCID: PMC8061710 (**Student Best Paper Nomination**)

Shehada ER, **He L**, Eikey EV, Jen M, Wong A, Young S, Zheng K. Characterizing frequent flyers of an emergency department using cluster analysis. In: Proceedings of the 2019 World Congress on Health and Biomedical Informatics (MEDINFO '19). 2019;158–61. PMID: 31437905

ABSTRACT OF THE DISSERTATION

Computational Analysis of Health Text

By

Lu He

Doctor of Philosophy in Informatics

University of California, Irvine, 2023

Professor Kai Zheng, FACMI, Chair

Health text ranging from patient-generated online forum posts to clinician-authored unstructured notes contain valuable information that can potentially improve healthcare service quality, patient experiences, and patient and population health outcomes. Health text data are also highly heterogeneous, produced in different contexts and serve different purposes, which require careful study design and methodological innovations to ensure study validity. However, the current practices of computational analysis on health text are often inconsistent and lack considerations of the contexts in which health text is produced.

My dissertation includes three major studies that analyzed different types of health text including public-generated social media data and clinical notes of patients with rare diseases. In the first study, I conducted a systematic literature review that revealed multiple issues in the current practices of how computational sentiment analysis is applied on health-related social media data. I also comprehensively evaluated the commonly used sentiment analysis tools on several social media datasets and found that they failed to accurately label the sentiments conveyed in health-related social media data. In the second study, I developed and applied computer-assisted qualitative analysis pipelines to analyze health-related social media data including tweets and online physician reviews. The results identified public attitudes and concerns toward mask wearing during the COVID-19 pandemic and patient

concerns around healthcare service quality. These insights contribute to better public health communication strategies and ways of enhancing patients' experiences when interacting with healthcare systems. In the third study, I switched gears to develop a pipeline that extracts various clinical entities including diagnosis, environmental exposures, substance use, performance status, and staging from unstructured notes of patients with lymphoid malignancies. The pipeline achieved satisfying performance and an error analysis identified issues with current documentation practices of key clinical information and provided recommendations for future improvement of the pipeline. The extracted clinical entities will be further used to facilitate clinical research to understand the association between environmental exposures and cancer outcomes.

Collectively, these studies contribute a set of methodological and empirical insights into how to design and choose an appropriate computational method to analyze different types of health text data. Moving forward, my future work will integrate and adapt the emerging Large Language Models into health text analysis, assess their performances, and identify potential biases when analyzing different types of health texts from various patient populations.

Chapter 1

Introduction

Health-related texts are generated and used in patients', or in fact, everyone's everyday life. When someone feels sick or uncomfortable, they may search on the internet and ask questions about their symptoms by posting on online forums. When patients see clinicians during clinical encounters, their information will be documented by clinicians in unstructured notes, which contain rich information about their symptoms, medical histories, medication prescriptions, and lab tests. When patients need clarification about lab test results, which are generally hard to interpret for laypeople, they may post again on online forums and receive answers from those who have similar conditions.

Health-related texts generated at all stages contain valuable information that can provide insights to address patient concerns, improve patient experiences and health outcomes, and facilitate clinical decision making and research. However, health-related texts are often of huge volume that are beyond the capacity of manual review. For example, in a study where we extracted the public's attitudes toward mask wearing in the United States during the pandemic, our final data included 771,268 tweets, which are not realistic for qualitative analysis [91].

Another challenge associated with analyzing health-related texts is their heterogeneity. Health-related texts encompass data generated by different stakeholders including patients, caregivers, clinicians, nurses, and the general public. In addition, the purposes of creating health-related texts can range from seeking information from the internet [162, 132], documenting medical history [35, 170], to expressing personal opinions regarding health-related events [91, 60, 69]. Therefore, the content and linguistic characteristics of health-related texts differ significantly, which necessitates customized computational methods for different types of health-related texts.

Using computational methods such as Natural Language Processing (NLP) to analyze large-scale, heterogeneous health-related texts has been extensively explored in the health informatics community. For example, sentiment analysis is widely used to extract opinions, attitudes, or concerns embedded in patients' and the public's posts on social media related to healthcare policies, personal protective equipment, and vaccination [60, 91, 69, 68]. Topic modeling, a technique that is used to automatically identify commonly occurring themes in textual data, is frequently applied on health-related texts to understand patient and public concerns for healthcare providers and policy makers to improve patient experiences and adjust policy making [92, 129, 199]. For text data generated in clinical settings such as clinical notes, the use of NLP techniques is well established to extract clinical entities and facilitate patient care, clinical decision support, and medical research [200, 21, 35, 63, 178, 126]. The field has reached maturity with several widely used NLP software such as the Clinical Language Annotation, Modeling, and Processing Toolkit (CLAMP) [180], clinical Text Analysis and Knowledge Extraction System (cTakes) [169], and STANZA [223].

1.1 Health text data: an overview

Health and healthcare is essentially driven by data. Everyday, extensive and diverse health data are generated, such as medical images, medical records, and lab test results, etc. Among them, textual data is an important resource that captures multiple aspects of patients' and the public's information about health. In this section, I will provide an overview of two distinct types of health text data that are generated by different stakeholders and in different settings, discuss the benefits and challenges of analyzing them to derive insights for health and healthcare research and practices, as well as how studies in this dissertation aim to fill the gaps.

1.1.1 The voices of patients and the public: health-related social media data

In the past decades, with the increased availability of Information and Communication Technologies (ICTs), patients, caregivers, and the public have unprecedented involvement in sharing and discussing health-related topics on public platforms. Both general social media platforms such as Twitter, Facebook, and Reddit and health-specific online forums such as MedHelp and BabyCenter are widely used by patients and the public. For example, it is estimated that MedHelp has more than 14 million users across the globe that exchange health-related information and seek support [135]. Patients and caregivers also self-organized groups on Facebook to share informational and emotional support with those who have similar health conditions such as Long COVID [168].

Social media data generated by patients, caregivers, and the public also serve as a valuable source for research during the COVID-19 pandemic [83]. Tsao et al. identified five strands of research that utilized social media data for COVID-related studies, including assessing public

opinions, identifying misinformation and disinformation related to COVID, and monitoring COVID cases and outbreaks [193]. For example, researchers exploited social media data to identify patient-reported symptoms of COVID and long COVID to facilitate clinical research [168, 28].

While health-related social media data serve as a great resource for health and healthcare research, they also pose significant challenges that need to be addressed to ensure the validity of computational analysis.

Why computational analysis of health-related social media data is challenging?

Because social media platforms attract diverse user populations [154] and afford dynamic discussions of a variety of health-related topics, social media data are often heterogeneous and even messy for research purposes. Different from data collected from surveys, interviews, or experiments that are specifically designed for research, social media data are usually generated without any prior research design considerations. Xu et al. defined such data as *organic data* and identified validity issues in using organic data such as social media data and digital trace data for research purposes [212]. Tufekci also identified several methodological issues in using social media data for research such as platform design as interfering factors and user behaviors that change over time [194]. Similarly, Olteanu et al. also reviewed the threats to research validity when using social data such as social media posts for research and proposed a paradigm to mitigate those issues [149].

While such reflections are emerging, the discussion of health-related social media data for research use is relatively sparse, despite numerous empirical studies that directly used social media data to derive insights for health and healthcare-related topics. Health-related social media data face distinct issues that may not be identical to the ones raised in the general social media domain [48]. In this section, I identify and discuss some of the challenges that

arose from using computational text analysis on health-related social media data.

Difficulties in reliably retrieving relevant health-related social media data

Due to the organic nature of social media data, retrieving relevant data that can answer research questions remains a fundamental challenge. For example, many health conditions are highly personal and even stigmatized such as eating disorders, depression, and sexually transmitted diseases (STDs). Therefore, social media discussions related to these health conditions often contain special vocabularies designed by users so that social media platforms cannot locate and censor them [48]. Locating relevant health-related social media data remains a challenging task due to the evolving and complex consumer-created vocabularies and platform design and policies. Kim et al. proposed and tested a data collection and filtering framework to assist reliable retrieval of health-related social media data for research [117]. However, Kim et al. also noted that developing keywords for data filtering is a tedious process, with the need to carefully monitor the precision and recall of each keyword and use human coding to assess their retrieval performance. Cummins suggested that keywords for searching relevant health-related social media discussions could be greatly enhanced by using word embedding techniques to capture relevant terms that may otherwise not identified by human coders [57]. For example, using the Word2Vec technique [138], Cummins identified additional keywords such as "vaxxed" and "jab" that are relevant to vaccination but due to their highly colloquial nature, they were not used as keywords in a previous study, which could result in a significant loss of relevant data for further analysis [128]. Similarly, Tong et al. also proposed using word embedding methods as well as network analysis to identify additional keywords based on human coders' initial keyword lists to broaden the inclusion of relevant health-related social media data [192].

Difficulties in filtering health-related social media data

Besides difficulties in reliably retrieving relevant health-related social media data, there remain several challenges in preprocessing the retrieved data for computational analysis. For

example, automated programs such as social bots can generate a huge number of social media posts in a short period of time to attract users and sometimes even manipulate public discussions [75, 24]. Social bots can "contaminate" health-related social media discussions, and including content contributed by social bots in further analysis may lead to incorrect and biased conclusions of public opinions. It is estimated that more than 70% of tweets related to e-cigarettes are potentially posted by social bots [54]. In addition, the content posted by social bots also differs from those posted by individual users. Allem et al. found that social bots are more likely to post content that support the use of e-cigarettes on Twitter [22]. To cope, Allem and Ferrara called for "debiasing" social media data when using them for health-related research to ensure the validity of study results [23].

Social bots are not the only actors that may contaminate social media-based health research. For studies that are interested in individuals' opinions and experiences related to health and healthcare, the presence of organizational accounts in collected data may also lead to biased results. For example, Zhang et al. found that organizational accounts post more chemotherapy-related content on Twitter than individual accounts, and these two types of accounts also show distinct characteristics such as their profile languages and interaction patterns [221]. Therefore, for researchers who plan to use tweets to study patients' and public's opinions toward chemotherapy, they should carefully filter out content from organizational accounts.

Difficulties in choosing and developing appropriate computational text analysis tools for health and social media contexts

The high heterogeneity of health-related social media data poses challenges for researchers to choose appropriate computational text analysis tools that can produce accurate and replicable results. Social media data collected from different platforms, related to different health-related issues, or even at different times can exhibit different linguistic characteristics because user populations and behaviors and platform designs differ and constantly evolve

[88, 165, 210]. For example, Roccetti et al. found that patients with Crohn's disease tend to discuss health-related issues more often on Facebook rather than on Twitter, which is potentially due to the sensitive nature of the health condition [165]. The selection of computational text analysis tools should therefore be driven by the characteristics of study data that is shaped by platform design, user behaviors, and the specific health topic studied. However, our systematic review on the use computational sentiment analysis on health-related social media data revealed that the selection of the tools are rather arbitrary in the health informatics community, often with little to no justification and pre-study evaluation [93]. This is problematic for the research community because the study results may be biased and incorrect, which could lead to inaccurate understanding of patients' and public perceptions toward health-related issues, and even incompatible health policies and mismatched health-related resource allocations.

Difficulties in validating social media-based research for health and healthcare

While many empirical studies that applied computational analysis on health-related social media data stated that social media data can complement or even replace traditional research data such as surveys, recent research has called for more caution and presented conflicting results derived from social media-based studies and surveys. For instance, Joseph et al. found discrepancies between self-reported stances collected through surveys toward issues such as masks and vaccines versus stances expressed in social media data by the same individuals [111]. The authors noted that measures based on labels from human annotators who coded the stances from the participants' social media data tended to underestimate the percentages of neutral stances. In addition, surveys and social media data collected at different time points may also measure different aspects of participants' stances toward health-related measures. These results call for more work to study whether and when are social media data reliable for measuring patients' and public's opinions toward health-related issues, and to what extent can we trust the results.

1.1.2 Rich but dense: clinical texts

Unstructured clinical notes such as admission notes, discharge notes, and radiology reports contain valuable information that serves many purposes such as documenting patient-provider interactions during clinical encounters, recording clinicians' judgements and rationales, providing evidence for billing purposes, and facilitating clinical research [166]. With the increased adoption of Electronic Health Records (EHR) systems in the United States and across the globe, clinical notes occupy a central role in clinicians' everyday work. Clinical notes have also been the main data source of research in the field of health informatics, in part due to the rich and multifaceted information captured in unstructured format.

While there are countless health informatics studies on clinical notes, a comprehensive review is out of the scope of this dissertation. In the following sections, I provide a focused overview on the use of Natural Language Processing (NLP) techniques to analyze clinical notes and identify gaps in the current practices.

Clinical Natural Language Processing: a brief overview

Information extraction from clinical notes

Information extraction (IE) refers to automatically extracting and encoding concepts, entities, and relations from text data [206, 56]. Common IE tasks include named entity extraction (NER) such as identifying mentions of locations, names, and time from texts [142] and relationship extraction that associates multiple entities (e.g., person and actions, medication and frequencies) [144]. These tasks are of great value for the clinical domain, because clinical entities such as diagnosis, medical history, and medication embedded in unstructured clinical notes can greatly assist clinicians' work and clinical research but are often tedious to extract manually. Therefore, clinical IE remains a popular field in health informatics that has seen continued methodological innovations, applications, and real-world implementations. Wang

et al. conducted a systematic review and identified 263 studies that applied IE on clinical notes from 2008 to 2016 [206]. They summarized the uses of clinical IE, including assisting patient care in many specialties such as chronic diseases and cancers, optimizing clinical workflow, and identifying drug-related information such as adverse drug events. They also noted that while machine learning-based IE systems achieve satisfying performances, the generalizability and portability of these systems remain limited.

More recently, many IE tasks are enhanced with the emerging pre-trained language models such as Bidirectional Encoder Representations from Transformers (BERT) [66]. Language models trained on large corpora can effectively capture semantic information and with fine tuning on study data, they can achieve even higher performances. For example, Si et al. evaluated traditional word embeddings and language models on various clinical IE tasks and found that using language models such as BERT that were pre-trained on domain specific corpora such as the Medical Information Mart for Intensive Care (MIMIC) dataset achieved the best performance on all IE tasks [174]. Further, Mulyar et al. developed the Multi-Clinical BERT model that can simultaneously perform multiple NLP tasks including NER and Relation Extraction at the same time [141]. The model achieved comparable though slightly degraded performances when compared to task-specific models. Multi-Clinical BERT even achieved slightly higher performance than the task-specific Clinical BERT model on extracting problems, treatments and tests.

Clinical entities and relations extracted from clinical notes using IE techniques can be further used in downstream tasks such as identifying patient cohorts and predicting patient outcomes. For example, traditional practices that only use structured data such as the International Classification of Diseases (ICD), lab tests, and medication history to identify patient cohorts often miss patients whose clinical information is not comprehensively captured by the structured codes. Applying IE to extract more comprehensive information from clinical notes can greatly improve the retrieval of patient cohorts [173]. Other downstream

applications using information extracted from IE tasks include patient outcome prediction, such as predicting patients' survival and length of stay [197]. For instance, Sterckx et al. developed an IE system to automatically extract information such as medication, tests, and blood loss conditions from clinical notes of pregnant patients [182]. The extracted information was then used to develop multiple machine learning models to predict preterm birth risks and found that the model using variables extracted through the IE system not only achieved high accuracy but also higher interpretability.

Challenges of developing and implementing clinical NLP systems

Despite the increased interests and successes in developing and applying clinical NLP and IE systems in the health and healthcare domain, several challenges remain.

First, acquiring high-quality annotations to train, fine-tune, and evaluate clinical NLP and IE systems is challenging, both due to the complexities and heterogeneity of how clinical information is documented in unstructured notes and the difficulties in recruiting clinician annotators. Various strategies have been proposed and tested to cope with this challenge. For example, active learning has been applied in various clinical NLP and IE tasks to automatically calculate and select notes that can maximally enhance the utility of model training and therefore decrease the number of notes clinicians need to annotate [126, 127, 50]. Data augmentation methods such as generating weak labels to reduce clinician annotators' burden have also been tested and achieved satisfying performance [205]. More recently, with the advances of Large Language Models (LLMs) such as ChatGPT and GPT3, researchers started to explore opportunities of using little to no annotated data to develop NLP and IE systems by providing curated prompts to LLMs such as ChatGPT [218, 208]. However, such explorations are still rather sparse in the health informatics field, which may be in part due to the restrictions of using commercial LLMs on clinical data that contain Protected Health Information (PHI) [155].

Second, even for clinical NLP and IE systems that achieve high performance on local data, it has been shown that they may have limited generalizability on out of distribution (OOD) data and low portability to other study sites. Sohn et al. found that semantic similarities in notes across institutions can lead to higher portability of clinical NLP systems in identifying asthma patients using clinical notes [177]. Mehrabi found that with customization to local study data, clinical NLP systems can further improve their performances in identifying patients with family history of pancreatic cancer [136].

Third, the integration of clinical NLP and IE systems into clinical decision support systems (CDSSs) and clinical workflows to assist clicians' work in real-world settings still faces significant challenges [63, 20]. While there is an increasing number of studies that developed state-of-the-art clinical NLP and IE systems on benchmark datasets, studies that reported experiences from implementing clinical NLP and IE systems in clinical settings are still relatively sparse, indicating a lack of real-world implementations and successful experiences. Lederman et al. proposed the paradigm of "task as needed", which states that the development of clinical NLP and IE systems should be closely connected with clinicians' needs and CDSSs, instaed of solely focuses on the technical performances of NLP models [120].

Fourth, while the development and application of clinical NLP and IE systems are well established and still see continued improvement for more prevalent diseases and health conditions, they are still underdeveloped for rare diseases [200]. This, again, may be due to the fact that rare disease clinicians are relatively difficult to recruit and the amount of clinical data available for patients with rare diseases is smaller and thus limits the ability to train and develop reliable models.

1.2 Contributions

My dissertation research makes multiple contributions to Health Informatics.

First, I systematically reviewed and synthesized the current practices of using computational methods on health-related social media data. I developed a checklist to assist researchers to conduct more standardized and rigorous computational analysis of health-related social media data. I evaluated commonly used computational tools and provided insights into how to tailor them for health-related contexts.

Second, I developed computer-assisted qualitative analysis methods to effeciently analyze patient and public-generated narratives on social media. I derived empirical insights from these large-scale health texts to understand patient and public concerns regarding healthcare service quality and public health crises.

Third, I developed an NLP pipeline that automatically extracts clinical information and social determinants of health from clinical notes of patients with lymphoid malignancies, a rare cancer. The results are further used to assist clinical decision making, facilitate clinical research, and understand documentation patterns and potential biases in the Veterans Affairs (VA) Health System.

1.3 Dissertation overview

This dissertation includes three main studies in six chapters.

Chapter 1 provides a brief overview of health text data including social media data and clinical texts and applications and challenges of using computational analysis on health text. The chapter also describes the contributions made by my dissertation research.

Chapter 2 presents the first study, in which I systematically reviewed studies that used computational sentiment analysis on health-related social media data. The study further evaluated commonly used computational tools and provided insights into how to tailor them for health-related social media data.

Chapter 3 describes the second study, in which I developed and applied computer-assisted qualitative analysis pipeline to analyze patient and public-generated social media data pertaining to their opinions toward health service quality and public health events.

Chapter 4 describes the third study, in which I developed and applied an NLP pipeline to extract clinical information and social determinants of health from clinical notes of patients with lymphoid malignancies.

Chapter 5 summarizes the empirical and methodological contributions made by my dissertation research. This chapter discusses the implications for computational analysis of health text in the era of Large Language Models (LLMs) and my plans for future research including incorporating LLMs and multimodal health data.

Chapter 6 concludes the dissertation.

Chapter 2

Study 1: Evaluating and improving the use of computational analysis of health-related social media data

2.1 Study 1A: Developing a standardized protocol for computational sentiment analysis research using health-related social media data

2.1.1 Study summary

Sentiment analysis has been widely used in the health informatics community to assess patients' and the public's opinions toward health-related issues such as vaccination and healthcare policy reforms. However, the selection and use of sentiment analysis tools is often arbitrary and inconsistent. In this study, I conducted a systematic literature review that

summarized and synthesized the research design and reporting practices of studies that used computational sentiment analysis on health-related social media data. The study revealed a high level of inconsistency of how social media platforms were selected for study, how data were collected and processed, and how computational sentiment analysis tools were selected, validated, and applied. I developed the Protocol of Analysis of senTiment in Health (PATH) based on the review, which encompasses comprehensive study design and reporting items for researchers to conduct more rigorous computational analysis of health-related social media data.

2.1.2 Introduction

Social media platforms such as Twitter and Facebook provide a public forum for anyone to create and disseminate content related to health, healthcare, or public health. For example, patients commonly share their disease journeys and exchange informational and emotional support with others who have similar conditions [156, 37]. Social media is also commonly used by the general public to voice their opinions on issues such as important health policies (e.g., the Affordable Care Act [60] and the lockdown orders due to the COVID-19 pandemic [190]) and controversial medical interventions and treatments (e.g., human papillomavirus vaccination [HPV] [69, 68] and the use of hydroxychloroquine for treating COVID-19 [172]). Because social media data are generally publicly available, relatively easy to obtain (e.g., through platform-provided application programming interface [API]), and are contributed by geographically and demographically diverse user populations [154], they have become an increasingly important source of information used by researchers to investigate a wide range of health-related topics. In fact, prior studies have demonstrated that public opinions expressed on social media platforms are highly correlated with poll results based on conventional surveys, confirming the feasibility of using such data for rigorous scientific research [60].

The sheer amount of user-generated social media data makes them difficult to manually analyze. Qualitative studies on small, selective samples preclude generalization to larger datasets. With the recent advances in natural language processing (NLP) and the increasing computing capability to process big data, researchers have now been able to use cutting-edge NLP techniques to efficiently analyze large volumes of free-text data with minimal manual effort. Sentiment analysis, in particular, has received increasing attention. Sentiment analysis, also referred to as opinion mining [151], is "the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions toward entities such as products, services, organizations, individuals, issues, events, topics, and their attributes."[122] A simple keyword search using "sentiment analysis" or "opinion mining" in PubMed yielded 348 papers; most were published in the recent ten years and the majority were based on computational methods. For instance, Davis et al. used NLP to study the general public's sentiments toward the Affordable Care Act; [60] and Huppertz and Otto developed a machine-learning model to analyze Facebook posts to assess patient opinions regarding their healthcare providers [105].

To date, numerous computational sentiment analysis methods (hereafter referred to as "sentiment analyzers") have been developed, ranging from lexicon-based dictionary lookups to machine-learning algorithms [188, 27, 133]. These methods have demonstrated satisfactory performance across many research domains; even though studies have commonly acknowledged the challenges to analyzing sentiments embedded in social media data due to their unique characteristics such as frequent use of short text, informal expressions and layperson terms for medical concepts, and special communication gimmicks such as hashtags and emojis [64, 183].

While computational sentiment analysis is an invaluable tool for understanding health-related opinions expressed on social media platforms, in our prior work, we noticed multiple issues in how existing studies were conducted and how their results were reported

[95]. For example, the keywords used to retrieve social media content often do not take into account the unique characteristics of consumer language used in social media posts; and some studies made rather arbitrary research design choices such as whether to filter out content contributed by non-laypersons (e.g., advocate groups and pharmaceutical companies), or whether to retain special types of data (e.g., images/videos, hashtags, emojis, and hyperlinks). Many also appeared to simply borrow existing sentiment analyzers developed in non-health domains (e.g., movie review) without validating their appropriateness for the particular study context, even though it has been repeatedly reported that the poor cross-domain transferability of sentiment analyzers could lead to inaccurate interpretations of data, or completely wrong conclusions [151, 31]. These issues may diminish the validity of the research. Indeed, in the literature, several studies have pointed out that they may have a significant impact on research results and conclusions. For example, a recent study found that organizational accounts posted more tweets expressing a positive attitude toward E-Cigarettes than individual users [134]. Similarly, another study found that organizational tweets, which comprise more than 70% of the tweets related to the side effects chemotherapies, tend to be more neutral, compared to tweets posted by individual users [221]. Social bots (i.e., computer programs that generate tweets automatically) exhibit similar behavior. For example, Allem et al. showed that social bots were more likely to post pro-cannabis tweets than non-bot users [22]. These findings suggest that the study design decision on whether to, or whether not to, differentiate social media content based on content contributors could lead to different findings and conclusions when conducting sentiment analysis research. Further, in our previous work [95], we evaluated three commonly used sentiment analyzers by applying them to two manually annotated social media health datasets. We found that all of these tools demonstrated poor performance, incorrectly classifying the neutrality of the posts in over 50% of the cases, compared to the sentiment labels assigned by human annotators. Further, inconsistencies in how different methods and tools were chosen and applied make it difficult to compare and synthesize results across studies, hindering our

ability to accumulate knowledge as a community. These observations motivated this work, through which we characterized common methodological and results reporting issues found in this body of literature, in order to develop a standardized protocol, which we refer to as the Protocol of Analysis of senTiment in Health (PATH), that may contribute to improving the quality and results comparability of future sentiment analysis research using health-related social media data, and other social media data analyses more broadly.

2.1.3 Objective

The objectives of this study were two-fold:
(1) to conduct a systematic review of the literature to identify common issues in research design and results reporting among studies that applied computational sentiment analysis to social media data on topics related to health, health care, or public health; and
(2) to develop the PATH based on the analysis of the relevant literature.

2.1.4 Material and methods

Systematic literature review

We conducted the search in January 2020 using 3 literature databases: PubMed, IEEE Xplore, and the ACM Digital Library. We included articles published in English and in peer-reviewed journals or conferences over a 10-year period between January 1, 2010, and December 31, 2019. Development of the search query (Table 2.1) was informed by previous literature reviews on the use of computational methods for analyzing health-related social media text [84, 48, 229, 217, 140]. We also supplemented our literature search results with articles referenced in these existing reviews.

((social media) OR (social network*) OR (social web*) OR (online social network*) OR (support group*) OR (Web 2.0) OR (Facebook) OR (Twitter) OR (MySpace) OR (Instagram) OR (YouTube) OR (Tumblr) OR (MedHelp) OR (WebMD) OR (online health communit*) OR (online forum*) OR (message board*) OR (discussion group*)) AND ((sentiment analysis) OR (opinion mining)) AND health*

Table 2.1: Search query

Following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guideline [139], we first screened titles and abstracts of the retrieved articles by applying the following exclusion criteria (1) studies conducted in topic areas not relevant to health, health care, or public health; (2) studies that analyzed non-English social media content; (3) studies that only performed manual review (eg, qualitative content analysis) of the data, as this study concerns sentiment analysis research that uses computational methods; and (4) studies that focused on development of new sentiment analyzers (eg, to report the algorithmic or mathematical under-pinning of a new sentiment analysis algorithm), or development of new software architectures (eg, to provide real-time sentiment analysis through cloud-based services), instead of analyzing social media data to generate empirical insights. Two authors (L.H. and Z.H.) independently screened the titles and abstracts of a random set of 50 articles. The screening results were discussed, and disagreements were resolved through consensus development research meetings. The remaining titles and abstracts were evenly split into 2 sets and separately reviewed. Then, full texts of the articles meeting the inclusion and exclusion criteria were retrieved and independently screened for eligibility by 2 authors (L.H. and T.Y.), who also independently extracted data from the final set of articles included in the review. Interrater reliability was assessed whenever applicable.

Development of the PATH protocol

We developed the PATH protocol through the following 3 steps. First, using the qualitative deductive coding and constant comparison approach [55], we identified and categorized distinctive research design choices that needed to be commonly made in relevant studies (eg, how to retrieve social media data and what sentiment analyzer to use). Then, we analyzed inconsistencies among the existing studies on these design choices and, when applicable, whether the rationale for a made choice was reported in the article. Finally, we synthesized the results from the analyses above to produce the PATH, the objective of which is to minimize such inconsistencies in order to improve the validity and results comparability of future sentiment analysis research in health.

2.1.5 Results

The PRISMA flow diagram exhibiting the screening process is reported in Figure 2.1. The literature search returned 417 results; 409 remained after duplicated entries were removed. The first round of screening based on titles and abstracts yielded 158 potentially relevant articles. The interrater agreement ratio was 0.88. Of these, 75 were deemed relevant upon a review of their full texts. The interrater agreement ratio was 0.94. We then conducted a citation analysis to identify additional relevant articles, which resulted in 14 more articles added. The final set selected for qualitative synthesis thus consisted of a total of 89 articles.

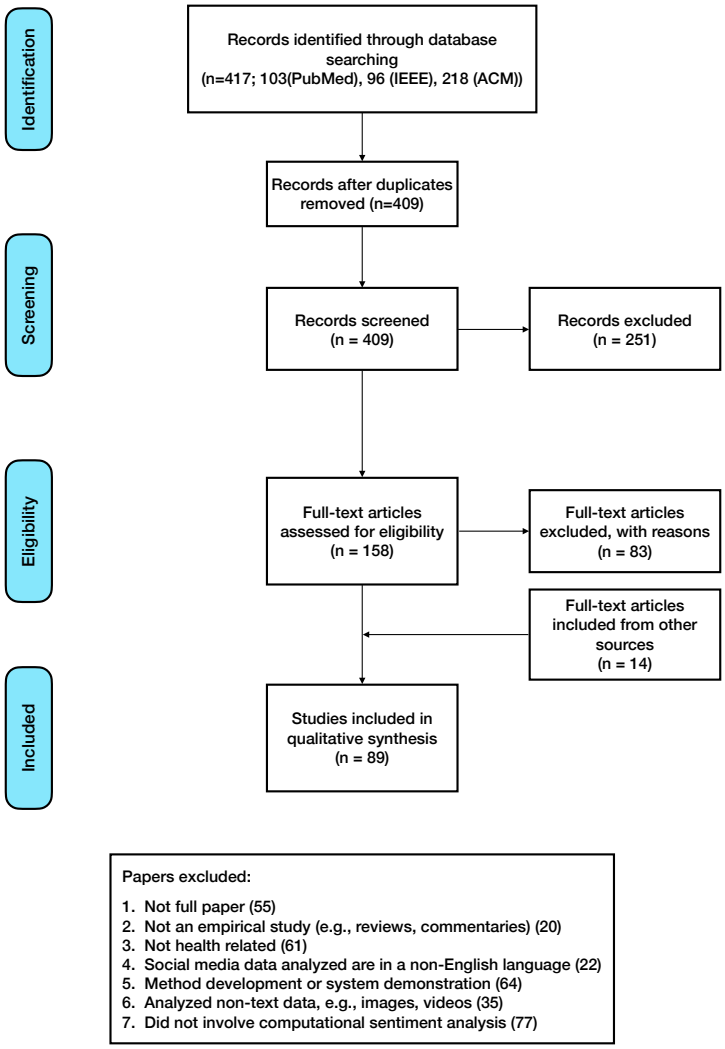

Figure 2.1: Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Flow Diagram.

Summary statistics of included papers

Of the 89 articles included in the review, most (n = 58) were published between 2017 and 2019. More than half (n = 51) analyzed Twitter data. The second and third most popularly studied platforms were Facebook (n = 5) and YouTube (n = 4), respectively. Besides these

general-purpose social media sites, some studies (n = 17) also examined health-specific online communities such as MedHelp (n = 3), CancerSurvivorNetwork (n = 3), JuiceDB (n = 2), Breast- Cancer.org (n = 2), WebMD (n = 1), QuitNet (n = 1), TalkLife (n = 1), LiveJournal (n = 1), Drug.com (n = 1), GLOBALink (n = 1), and BecomeAnEX.org (n = 1).

Research design choices

Based on the articles reviewed, we first identified a list of distinctive research design choices that needed to be commonly made in conducting health-related computational sentiment analysis research using social media data. We then organized these design choices, reported in Table 2.2, according to the following 3 dimensions: (1) platform selection, (2) data curation; and (3) sentiment analysis method. The first dimension concerns how studies choose the appropriate social media platform that would be most informative for the research questions at hand; the second dimension concerns how to retrieve and curate relevant data that do not introduce unwanted biases (eg, whether to retain or remove advertisements posted by pharmaceutical companies), or loss of critical information (eg, whether to retain, remove, or substitute hashtags and emojis). The third dimension concerns how to select the appropriate analytical tool best suited for the particular study context, and whether and how to validate the tool before applying it to the study data.

As shown in Table 2.2, a number of studies (n = 24 of 63 applicable) did not use, or did not report, any method for determining the relevance of the research data retrieved. Most (n = 45 of 58 applicable) did not differentiate the data based on content creator. While about one-third of the studies (n = 32 of 89) reported how special types of data such as hashtags and emojis were handled (ie, retain, remove, or substitute), less than half (n = 12) provided a rationale for the choice made. For the last dimension, sentiment analysis method, the majority of the studies (n = 49 of 89) did not provide any justification as to why the particular sentiment analyzer or the machine learning model was chosen. Additionally, 53

studies of 89 did not validate the sentiment analyzer for their particular study context. Among those that did use manually annotated data for sentiment analyzer validation, many (n = 11 of 26 applicable) did not involve multiple coders. Last, among the studies that used machine learning (n = 29), 10 did not describe the features selected. Among those that did, several (n = 7 of 19) did not justify the feature selection process.

Dimension	Design Choices	Description	Example
Platform selection	Which social media platform provides data that are most informative to answer the research questions of the study?	As different social media platforms attract different types of users and foster different forms of communication, studies may want to evaluate available options and decide which one(s) would provide the best information for studying the research topic of interest.	"WebMD.com hosts one of the few online communities that offer moderators in patient forums. Their diabetes community shows the most active participation of both patients and moderators among other WebMD communities [104]."
Data curation	What is the strategy used to retrieve relevant data?	Procedures for identifying potentially relevant social media posts based on keywords, user characteristics, or other means of information retrieval; and procedures for determining the relevance and comprehensiveness of the data rerieved.	"We started with a set of relevant seed keywords (eg, 'lynch syndrome'). Then, we searched on Twitter with these keywords to retrieve a sample of tweets, evaluated whether the retrieved tweets were indeed relevant to Lynch syndrome, and identified additional keywords to be used for the next rounds of searches [30]."
	Whether to differentiate data based on content creator?	Social media data can be contributed by different entities such as laypersons, healthcare providers, health systems, government agencies, advocacy groups, and pharmaceutical companies. Depending on the research objective, studies may want to treat data differently based on the creator of the content.	"In order to gain insights into the opinion and experience of cancer patients about chemo-therapy, these cancer-related user accounts were classified into two groups: individual accounts and organization accounts. The individual accounts belonged to cancer patients as well as their families, whereas the organization accounts include organizations, oncologists, news sources, and personnel who are neither patients nor family members [221]."

Table 2.2 continued from previous page

Dimension	Design Choices	Description	Example
	How to handle special types of data.	As social media data frequently contain elements such as images or videos, hashtags, emojis, and hyperlinks, studies should determine whether to retain, remove, or substitute such data at the preprocessing stage, and explain the rationale for the approach chosen and its implications for study results.	"We cleaned out contents such as emoji icons, urls, from each tweet. By observing the data, we noticed that hashtags tended to store very important content. For instance, a lot of the anti-vaccine tweets contained '#CDCwhistleblower'. Therefore, instead of deleting the content of hashtags, we only deleted the '#' symbols and used the hashtag content as part of the content of tweets to train the models [219]."
Sentiment analysis method	Which sentiment analyzer is most suited for the study context, particularly the characteristics of the social media data to be analyzed?	Among many options available, which sentiment analyzer to choose that would maximize the quality of the study analysis.	In this study, we use SentiStrength as (i) it has been used to measure the emotional content in online ED communities and shown good inter-rater reliability; (ii) it is designed for short informal texts with abbreviations and slang, and thus suitable to process tweets [201]."
	Whether to validate the selected sentiment analyzer on the study data.	Even if the selected sentiment analyzer has been applied by others to similar datasets in the past, it may still be worthwhile to conduct prestudy validation to ensure it performs satisfactorily on the data collected for the particular study.	"In addition to the already mentioned evaluation of the accuracy and performance of EMOTIVE, a brief qualitative manual review of a sample of EMOTIVE's output showed a consistent and correctly categorized set of emotions among the seven basic emotions [86]."
	If prestudy validation is to be performed, whether to obtain a manually annotated dataset for training or evaluation purposes.	To validate the performance of the selected sentiment analyzer, studies may want to obtain manual annotations of a subset of the study data, ideally with multiple coders so that interrater reliability can be assessed.	"To identify and calibrate the classification model, 298 randomly selected posts were manually labeled by two independent annotators as belonging to either the positive or negative sentiment class. Cohen's k statistics (k=0.82) suggested high inter-annotator agreement. Then the two annotators discussed posts whose sentiment they initially disagreed on until they reached a consensus on sentiment labels [224]."

Table 2.2 continued from previous page

Dimension	Design Choices	Description	Example
	If prestudy validation is performed, whether the validation results are computed and reported using established quantitative metrics.	Studies should report the validation results based on commonly used quantitative evaluation metrics such as F score, or receiver-operating characteristic curve.	"For this dataset, classifiers performed reasonably well, with F1 scores ranging from 0.48 to 0.68. However, the logistic regression classifier used with the n-gram model performed the best with an F1 score of 0.68. This performance is comparable with that in similar studies [131]."
	Design choices specifically related to developing or training machine learning–based models.	In developing or training machine learning–based sentiment analyzers, studies should evaluate different competing models (eg, support vector machine, decision trees), as well as different features that may be selected to train the model (eg, bag of words, word vectors).	"The n-gram model performed slightly better than the word-embedding model. For this dataset, classifiers performed reasonably well, with F1 scores ranging from 0.48 to 0.68 [131]."

Table 2.2: Distinctive research design choices

Methodological and reporting inconsistencies among the existing studies

Next, we assessed inconsistencies in how the existing studies reviewed made the aforementioned research design choices, and how they reported the rationale of making such choices, or the lack thereof. The results are shown in Table 2.3.

Dimension	Item	Reported	Not Reported	Not applicable
Platform Selection (PS)	PS1. Description of the social media platform studied	89	0	0
	PS1-A. Justifications for selecting the social media platform	79	10	0
Data Curation (DC)	DC1. Methods for retrieving study data	85	4	0
	DC2. Methods for determining the relevance and comprehensiveness of the data retrieved	39	24	26
	DC3. Differentiated treatment based on content creator	13	45	31
	DC4. Handling of special types of data (e.g., images/videos, hashtags, emojis, hyperlinks)	32	57	0
	DC4-A. Justifications for how special types of data are handled (N=32)	12	20	0
Sentiment Analysis (SAM)	SAM1. Description of the sentiment analyzer used	83	6	0
	SAM1-A. Justifications for selecting the sentiment analyzer	40	49	0
	SAM2. (If machine learning) Description of the features selected (N=29)	19	10	0
	SAM2-A. (If machine learning) Justifications for selecting the features (N=19)	12	7	0
	SAM3. Validation of the sentiment analyzer before applying it to study data	36	53	0
	SAM3-A. Annotated data used for validation or training (N=36)	30	6	0
	SAM3-B. Whether multiple coders were involved in independently annotating the data (N=30)	15	11	4
	SAM3-B-1. If multiple coders were involved, whether inter-rater reliability was quantitatively assessed and reported (N=15)	13	2	0

	SMA3-C. Use of quantitative evaluation metrics for reporting the validation results (N=36)	32	4	0

Table 2.3: Methodological and reporting inconsistencies among the existing studies (N = 89)

PROTOCOL OF ANALYSIS OF SENTIMENT IN HEALTH (PATH)

	Design or Reporting Considerations	Description
Platform Selection	☐ **PS1. Description of the social media platform studied**	Characteristics of the social media platform studied such as intended audience and interaction modality.
	☐ **PS1-A. Justifications for selecting the social media platform**	Why is the chosen social media platform provide data that are most informative to answer the research questions of the study?
Data Curation	☐ **DC1. Methods for retrieving study data**	What is the strategy used to retrieve study data, e.g., by keywords search or by targeting particular users with certain characteristics?
	☐ **DC2. Methods for determining the relevance, and comprehensiveness (if applicable), of the data retrieved**	What are the methods used to ensure that the data retrieved are pertinent to the research topic(s) of interest, and are reasonably compete?
	☐ **DC3. Differentiated treatment based on content creator**	Are data contributed by different entities such as laypersons, healthcare providers, and pharmaceutical companies treated differently?
	☐ **DC4. Handling of special types of data (e.g., images/videos, hashtags, emojis, hyperlinks)**	Are special types of data retained, removed, or substituted in the analysis?
	☐ **DC4-A. Justifications for how special types of data are handled**	Why are special types of data handled in the particular way, and what are the implications?
Sentiment Analysis Methods	☐ **SAM1. Description of the sentiment analyzer**	What is the sentiment analysis tool or machine-learning model used in the study?
	☐ **SAM1-A. Justifications for selecting the sentiment analyzer**	Why is the chosen sentiment analyzer most suited for the study context?
	☐ **SAM2. (If machine learning) Description of the features selected**	What are the features used in the machine-learning model, and how are they selected?
	☐ **SAM2-A. (If machine learning) Justifications for selecting the features**	Why are the chosen features most suited for the study context?
	☐ **SAM3. Validation of the sentiment analyzer before applying it to study data**	How is the performance of the chosen sentiment analyzer assessed against the study data?
	☐ **SAM3-A. Annotated data used for validation or training**	What are the training or evaluation data used, and how are these data obtained?
	☐ **SAM3-B. Whether multiple coders were involved in independently annotating the data**	If applicable, are multiple coders involved in independently annotating the training or evaluation data?
	☐ **SAM3-B-1. If multiple coders were involved, whether inter-rater reliability was quantitively assessed and reported**	What is the quantitative inter-rater reliability between the multiple coders?
	☐ **SMA3-C. Use of quantitative evaluation metrics for reporting validation results**	What are the quantitative metrics (e.g., F-score, ROC) used to assess the analyzer performance?

Figure 2.2: PATH

2.1.6 Discussion

Social media has become an important resource of information for researchers to better understand patient journeys, their interactions with health systems and healthcare providers, as well as patients' and the general public's opinion toward important health policies and controversial medical interventions and treatments. A large number of such studies have been published in recent years, most of which used computational methods to analyze the sentiments expressed in the data. However, based on our systematic analysis of the literature, we found that there is a substantial degree of inconsistency in how such studies were conducted and how their results were reported, which may diminish the quality of research in addition to making it difficult to conduct meta-analyses to accumulate generalizable knowledge as a field. Subsequently, we discuss some of these methodological or reporting inconsistencies identified through this work and how they may affect research validity and comparability of results across studies.

First, some studies did not at all describe the process of sifting through available social media platforms to choose the ones that were most informative, in comparison with other competing social media outlets, to best answer the research questions at hand. Many simply stated that the chosen platform was a popularly used one, or commonly studied in prior research, or provided the easiest access to data. We believe such justifications, while may be reasonable due to practical reasons (eg, difficulties in accessing patient conversations in private Facebook groups), could potentially threaten the validity of the study, and researchers should use all means necessary to minimize possible data biases and improve the generalizability of their research results and conclusions. Indeed, previous studies that compared multiple social media platforms did find that different venues afforded different health content [165, 210], appealed to different user populations with distinctive characteristics [88], or featured different interaction modality (eg, moderated vs not moderated) that may affect the nature of the discourses [210]. All of these factors could have significant implications on the results

and conclusions of sentiment analysis research using health-related social media data.

Second, many existing studies did not conduct, or did not report, the data curation process for determining the relevance and comprehensiveness, if applicable, of the study data. This is particularly concerning in the analysis of health-related social media content because of the frequent use of ambiguous acronyms and abbreviations (eg, SOB for shortness of breath), similar medical concepts that may not be generally differentiated by laypersons (eg, dementia and Alzheimer's disease), and mixed usage of consumer language vs professional terms (eg, heart attack vs myocardial infarction). Further, very few studies treated their study data differently based on content creator, being laypersons, healthcare providers, health systems, government agencies, advocacy groups, or pharmaceutical companies. Depending on the objective of the study, this could result in "contaminated" data that did not truly reflect the sentiments of the target study population, and could consequently lead to imprecise or incorrect conclusions [221, 22, 23, 34]. Future studies may consider adopting the methods proposed by Kim et al. [117] and Adams et al. [18] on how to develop and iteratively refine search keywords (eg, through word embeddings) for retrieving the content of interest from social media platforms, and how to thoroughly evaluate the relevance, and comprehensiveness (if applicable), of the information retrieved using manually annotated data. Furthermore, few studies described how they handled special types of data such as images or videos, hashtags, emojis, and hyperlinks, which are commonly used in social media discourses and can in fact convey important information about the sentiments being expressed [95, 100]. However, this process was omitted from most existing studies, or was only causally mentioned (eg, all special types of data were removed) without providing any justification as to how the particular handling method used might affect the study results.

Third, most studies did not provide a rationale for choosing among many different sentiment analyzers available. Only a small number of the studies validated the selected tool to assess its performance (ie, precision and recall) against the study data. This can be prob-

lematic, as prior research has repeatedly demonstrated that different sentiment analyzers, especially those general-purpose ones developed or trained on datasets from non-health domains (eg, movie reviews), could produce substantially different results due to their poor domain transferability and the idiosyncrasies of health-related social media conversations [95, 124]. Among the studies that did perform validation, only half involved multiple coders to annotate the training or evaluation data. This could also raise questions into research validity because previous studies have shown that annotating social media sentiments in general, and of health-related content in particular, is a challenging task even among experienced domain experts [58, 25]. Therefore, having a single pair of eyes would not be considered sufficient for assuring the quality of annotations of such data.

The research design and reporting recommendation that we developed through this study, PATH, represents an initial step to ad- dress each of these issues. Applying a standardized protocol such as PATH in future health-related social media sentiment analysis re- search may also produce a higher level of consistencies in research design, conduct, and reporting, leading toward better comparability of results across studies. We believe that some elements of PATH, such as platform selection, data curation, and tool validation, also apply broadly for other studies that use computational methods to analyze health-related social media content, beyond just sentiment analysis. Therefore, we hope that this work will stimulate more critical reflection and development of standardized research protocols in a broader scope of computational analysis of social media data.

This study has several limitations. While sentiment analysis is an important tool for analyzing social media data, other methods such as topic modeling, spatiotemporal analysis, and social network analysis are also popularly used, which are not addressed in this study. Further, while we hope all elements proposed in the PATH protocol should be adhered to in future relevant studies, we understand that some desired research design choices may not be attainable due to resource constraints (eg, cost-prohibitive to involve multiple coders to

annotate training or evaluation data) or practical reasons (eg, impossible to get data from the social media platform that provides the ideal user mix and the ideal content). The PATH protocol should therefore be interpreted as a set of recommended steps rather than mandatory requirements.

2.1.7 Conclusion

In this study, we systematically analyzed the body of literature that applied computational sentiment analysis to studying health-related social media content. The results highlighted a substantial degree of inconsistencies in how existing studies were conducted or how their results were reported. These findings led to the development of a recommended research design and reporting guideline, PATH. We believe that application of PATH in future sentiment analysis studies could lead to better research validity and comparability of results. The elements in the PATH protocol may also provide insights more broadly into other genres of research studies that use computational methods to analyze health-related social media data.

2.2 Study 1B: Empirical evaluation of computational sentiment analysis tools on health-related social media data

2.2.1 Study summary

While sentiment analysis is widely used to assess public and patients' opinions toward health-related matters, whether existing tools can produce reliable results on health-related social media data remains unknown. In this study, I conducted a comprehensive evaluation of eleven commonly used sentiment analysis tools on five health-related social media datasets, including Human Papillomavirus Vaccine, Health Care Reform, COVID-19 Masking, Vitals.com Physician Reviews, and the Breast Cancer Forum from MedHelp.org. I also conducted a qualitative error analysis to identify common sources of errors made by these tools. The results show that all tools performed poorly with an average weighted F1 score below 0.6. Further, the tools do not agree with each, with an average Fleiss Kappa score of 0.066. By comparing the sentiment classifications produced by the tools to the labels annotated by human coders, I found that applying sentiment anlaysis tools with low performance on health-related social media data may lead to significant overestimation or underestimation of certain sentiment categories, which could bias the interpretation of public perceptions. To provide insights for future sentiment analysis tool development, I identified two major causes for misclassification: (1) correct sentiment but on wrong subject(s) and (2) failure to properly interpret inexplicit/indirect sentiment expressions. This study warned that researchers should not blindly trust sentiment analysis tools even if they have been validated on other study data. The current tools are not adequate in accurately identifying the sentiments in health-related social media data and need significant customization and improvement to adapt them to the health and social media contexts.

2.2.2 Introduction

In recent years, there has been a proliferation of research involving analysis of user-generated data on social media platforms such as Facebook, Twitter, Instagram, and YouTube. As of October 2021, more than 20,000 PubMed-indexed papers reported studies that used social media data to understand a variety of research topics from patient journeys to how the public reacted to health policies and public health crises. Collectively, these studies have generated a wealth of knowledge on how individuals exchange information, opinions, or social and emotional support on the internet. These studies have also produced many novel natural language processing (NLP)-based analytical approaches for deriving valuable insights from large quantities of user-generated online text.

Sentiment analysis, "*the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions toward entities such as products, services, organizations, individuals, issues, events, topics, and their attitudes*" [122], is a commonly used NLP method for analyzing social media data. According to a recent systematic review [93], between 2010 and 2020, 89 papers applied this method in studies related to medical treatments (e.g., human papillomavirus [HPV] vaccination) [68], health policies (e.g., the Affordable Care Act) [60], and consumer satisfaction of healthcare services (e.g., using data from physician rating websites) [92]. Even though deep learning-based sentiment analysis methods, with models trained on the specific study data, have achieved remarkable performance, they require abundant manually annotated training data, extensive computational resources, and adequate technical expertise [222]. On the contrary, off-the-shelf sentiment analysis tools (both rule-based and pre-trained machine learning models) are easier to use and require little to no manually annotated training data and technical expertise. Indeed, a majority of the studies in the systematic literature review used off-the-shelf tools originally developed in non-health domains, such as Stanford NLP Sentiment pre-trained on movie reviews from IMDB [133] and the Hu & Liu Sentiment Lexicon extracted from product re-

views from Amazon.com [102]. While these tools demonstrated satisfactory performance in their original validation studies [102, 133], in most cases, they were used as a "black box" to analyze health-related social media data without proper evaluation and adaptation (e.g., model re-training) for the specific study context [93]. Thus, it remains unclear if the performance of these tools is sufficient enough to produce accurate sentiment classifications on health-related social media data, such as the general public's attitude toward mask or vaccination mandates during the COVID-19 pandemic, to inform important policy and public health decision making.

While this issue has been previously noted in discrete studies [58, 124, 95], to the best of our knowledge, no research to date has attempted to systematically evaluate the existing sentiment analysis tools in the context of analyzing health-related social media data. Further, no research has explored how suboptimal sentiment classification performance may impact the analysis and results interpretation of relevant empirical studies. As a matter of fact, many recently published studies continued to use off-the-shelf sentiment analysis tools without assessing their validity in the particular research context in which they were applied [79, 137, 73, 220, 118]. This paper aims to fill this gap. We comparatively assessed the performance of 11 commonly used sentiment analysis tools on seven social media datasets (five health-related and two non-health related) to investigate their validity and reliability. We then conducted a qualitative error analysis to examine the instances for which all sentiment analysis tools failed to produce accurate sentiment classification. The results may help to raise awareness among the research community regarding the limitation of the off-the-shelf sentiment analysis tools when they are applied to analyze social media data outside the domain in which they were originally developed. The results may also offer insights into future improvements of such tools.

2.2.3 Materials and methods

Sentiment analysis tools evaluated

Most of the existing off-the-shelf sentiment analysis tools can be classified as either sentiment lexicons or sentiment analyzers. A sentiment lexicon is a dictionary of words associated with sentiment labels (e. g., "brilliant" → positive vs. "terrible" → negative). Examples include Linguistic Inquiry and Word Count (LIWC) [188] and SentiWordNet [72]. Some sentiment lexicons are polarity-based, which only provide categorical labels (i.e., positive vs. negative vs. neutral); some compute a numeric score to indicate both polarity and intensity (e.g., 5 is more positive than 1). On the other hand, a sentiment analyzer is a packaged software that uses manually curated rules (e.g., VADER [107] such as incorporating punctuations (e.g., exclamation marks) in calculating sentiment intensity or machine-learning models (e.g., Stanford NLP [133]) trained on datasets annotated by human annotators.

In this study, we evaluated a total of 11 sentiment analysis tools listed in Tables 2.4 and 2.5. Seven of them are sentiment lexicons and four are sentiment analyzers. The selection of these tools was based on a prior systematic literature review in which we identified 19 relevant sentiment analysis tools [93]. For this study, we applied two additional selection criteria: (1) the tool must have been used in at least two different studies conducted by different research teams, and (2) the tool must be readily accessible, either for free or for a license fee. With these additional criteria, five tools were excluded because they were developed and validated in a single study and were not subsequently adopted by other researchers; and three were excluded because they were no longer available.

To compute sentiment polarity or scores using a sentiment lexicon, we followed the instructions provided in the manual. If no instructions are provided (i.e., ANEW and General Inquirer), we used the common practice in the literature, i.e., determining polarity based

Name	Size	Method of Development	Output	Sample Studies
Affective Norms for English Words (ANEW) [33] (free)	3,188	Manually curated from generic tweets	Polarity and intensity (-5 to 5)	Nguyen et al., [147] Ricard et al. [163]
AFINN-111 [108] (free)	2,477	Manually curated from generic tweets and the Urban Dictionary [16]	Polarity and intensity (-5 to 5)	Yang and Kuo, [213] Tighe et al. [191]
General Inquirer [3] (free)	3,626	Aggregated from multiple dictionaries, such as the Lasswell Value Dictionary [11]	Polarity (positive vs. negative)	Carrillo-de-Albornoz et al., [40] Oh et al [148]
Hu & Liu [102] (free)	6,789	Opinion words extracted from Amazon.com product reviews	Polarity (positive vs. negative)	Card et al.,[39] Li et al. [121]
LabMT [67] (free)	10,222	Manually curated from generic tweets, Google Books, music lyrics , and New York Times articles	Intensity (1 to 9)	Davis et al., [60] Chopan et al. [52]
LIWC-2015 [188] (one-time license fee of $89.5)	907	Manually curated from personal diaries and journals with expressions of emotions	Polarity (positive vs. negative)	Shen et al., [171] Oscar et al. [150]
SentiWordNet [72]	117,660	Opinion words generated from WordNet, an English lexical database [17]	Polarity and intensity (-5 to 5)	Rastegar-Moiarad et al.,[158] Wiley et al. [210]

Table 2.4: The seven sentiment lexicons evaluated.

Name	Type	Method of Development	Output	Sample Studies
VADER [107] (free)	Rule-based	Manually curated rules from analyzing a Twitter dataset	Polarity and intensity (-1 to 1)	Pawsey et al., [153] Pérez-Pérez et al.[157]
SentiStrength [189] (free for academic use)	Machine learning	Machine learning-based models trained using the vocabularies from LIWC and General Inquirer	Polarity and intensity (-5 to 5)	Gabarron et al., [80] Wang et al.[201]
Stanford NLP [133] (free)	Machine learning	Recurrent network trained on movie reviews	Polarity (positive, negative, neutral)	Haghighi et al.,[62] Talpada et al.[186]
TextBlob [13] (free)	Unspecified	Unspecified	Polarity and intensity (-1 to 1)	Luo et al.,[125] Zhang et al.[221]

Table 2.5: The four sentiment analyzers evaluated.

on counting the frequency of positive and negative words [102]. Similarly, for sentiment analyzers, we followed the instructions provided with the tools scores to discrete sentiment categories. If no instructions are provided (i.e., TextBlob), we used the common practice; for example, if a social media post's sentiment score was lower than the median, we classified it as negative.

Validation social media datasets

The validation social media datasets were identified through a systematic review of the literature that we previously conducted [93]. In addition, we added several new datasets that became available after the systematic review was published [92, 91]. Using this method, we arrived at a total of 17 candidate datasets. Of them, four are publicly available. For the other 13, we approached the authors to inquire if they were willing to share the annotated data for this study. Only one team agreed to provide us the data; the rest were either nonresponsive or declined due to IRB restrictions or other logistical reasons. Thus, the final number of health-related social media datasets included in this study is five. Three of them are Twitter-based: Health Care Reform ("HCR") [181], Human Papillomavirus ("HPV") Vaccine [68], and COVID-19 Masking ("Mask") [91]. The other two are based on the Breast Cancer Forum from MedHelp.org ("MedHelp") [203, 202] and Vitals.com Physician Reviews ("Vitals") [92].

For comparison, we also included two generic social media datasets that are not related to health, healthcare, or public health. These include (1) the IMDB Dataset ("IMDB") curated by Maas et al. from movie reviews [130], and (2) Sem-Eval-2016 Twitter Data ("Sem-Eval"), which has been used in a recurring NLP competition organized by the International Workshop on Semantic Evaluation [12]. These two datasets were commonly used by the machine-learning community to train and/ or evaluate machine learning-based sentiment analysis algorithms [29, 143]. Table 2.6 presents more details about these five health-related datasets and two generic datasets.

Dataset	Type	Original development and validation study	Size	Sentiment Classification
HCR	Health-related	Speriosu et al. [181]	2,315 words	Positive (701) Negative (1,034) Neutral (580)
HPV	Health-related	Du et al. [68]	3,093 tweets	Positive (964) Negative (1,044) Neutral (1,085)
Mask	Health-related	He et al. [91]	609 tweets	Positive (530) Negative (79)
MedHelp	Health-related	Wang et al. [202]	500 posts	Positive (63) Negative (239) Neutral (198)
Vitals	Health-related	He et al. [92]	50,000 reviews	Positive (25,000) Negative (25,000)
IMDB	Generic	Maas et al. [130]	50,000 reviews	Positive (25,000) Negative (25,000)
Sem-Eval	Generic	Nakov et al. [143]	28,466 tweets	Positive (10,905) Negative (4,472) Neutral (13,089)

Table 2.6: Description of the validation social media datasets.

Evaluation metrics

We used weighted F1 score to evaluate the overall performance of the sentiment analysis tools, calculated as: of social media posts classified into each of the three sentiment cate- gories by human annotators. We used weighted F1 score in this study because the distribution of sentiment polarity is often imbalanced in a dataset. To assess the reliability among the tools (i.e., consistency of the classification results produced by different tools), we used Fleiss' Kappa agreement ratio [76].

Qualitative error analysis

Next, we conducted a qualitative error analysis of the data that were consistently misclassified by all 11 tools. Two authors (LH and TY) conducted independent qualitative coding of a random sample of these instances using the constant comparison method [70] to identify com- mon reasons for misclassification. The analysis stopped when saturation was achieved [55]. Differences between the two reviewers were discussed and resolved in weekly research meetings.

2.2.4 Results

Overall performance of sentiment classification

Figure 2.3 presents the results of sentiment classification performance evaluation. From left to right, the 11 tools are ordered by the average weighed F1 scores that they achieved across the seven datasets. Overall, none of the tools produced satisfactory results. The average weighted F1 scores of all tools, on all datasets, are below 0.6. Three tools in particular—General Inquirer, SentiWordNet, and LabMT—produced an average weighted F1 score

ranging from 0.14 to 0.2. In other words, the accuracy of sentiment classification generated by these tools is much worse than by tossing a coin. Further, across the 11 tools, there is a great degree of variation in performance. For example, on the Vitals dataset, VADER achieved a weighted F1 score of 0.8, yet LabMT only achieved 0.1.

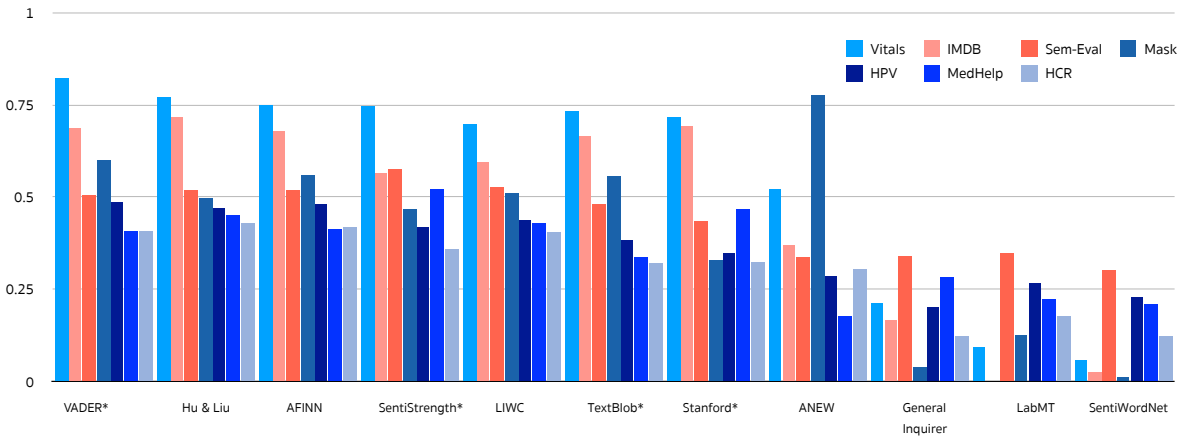

Figure 2.3: Weighted F1 scores.

The nature of a dataset appears to have a substantial impact on sentiment classification results. Among the tools that performed relatively better (average weighted F1 scores $>$ 0.5), their performance on the Vitals dataset is consistently the best, followed by the two generic datasets (IMDB and Sem-Eval), and then the Mask dataset. The only exception

is ANEW, which performed best on the Mask dataset (weighted F1 score: 0.78), followed by Vitals (weighted F1 score: 0.52). All tools performed poorly on the other three health datasets (HCR, HPV, and MedHelp), with most of them achieving a weighted F1 score lower than 0.4. Again, this performance is worse than that of tossing a coin.

Further, we did a drill-down analysis by grouping datasets with similar natures. First, most tools, except for ANEW, performed worse on tweets than on other texts. This pattern is however not observed for MedHelp, which contains longer texts. Furthermore, as shown in Fig. 1, the sentiment analyzers (denoted with an asterisk on the x-axis) do not appear to outperform the lexicon-based tools. In fact, the rule-based sentiment analyzer VADER produced better results than the machine learning-based analyzers (i.e., Stanford and SentiStrength). Next, we calculated the average weighted-F1 scores grouped by tool and data types, reported in Table 2.7. The results show that sentiment analyzers performed better than sentiment lexicons both on tweets and non-Twitter data. Both sentiment analyzers and lexicons performed better on non-tweets.

Overestimation and underestimation

Figure 2.4 presents a heatmap visualization exhibiting the degree to which the tools overestimated or underestimated positive (top left subgraph), negative (top right), or neutral (bottom left) sentiments. The color spectrum represents overestimation (red) or underestimation (blue). The numeric values in the cells are overestimation or underestimation ratios, calculated as the number of social media posts classified by machine divided by the number classified by human annotators, with underestimation denoted as negative values. For example, 6.6 in the top left subgraph means that ANEW, when applied to MedHelp, overestimated the positivity of the data by 6.6 folds; and -176.7 means that SentiWordNet underestimated the positivity of the Mask dataset by 176.7 folds.

The most salient pattern from Figure 2.4 is that all tools have a general tendency to overestimate the neutral category. Three tools–SentiWordNet, LabMT, and General Inquire–consistently underestimated both positive and negative categories, while consistently overestimating neutral. ANEW is again an outlier, which consistently overestimated positivity and underestimated negativity and neutrality. Across the datasets, the positivity of MedHelp was overestimated by most tools, while its negativity was underestimated. Similarly, the negativity of the Mask dataset was consistently overestimated, and the negativity of HCR and Vitals was consistently underestimated.

Figure 2.4 also shows a drastic level of variation across the tools. For example, for the MedHelp dataset, the positivity ranged from underestimation by over 60 folds (LabMT and SentiWordNet) to overestimation by more than 5 folds (TextBlob and ANEW). Similarly, for the Mask dataset, the negativity ranged from 40-fold underestimation to 2.8-fold overestimation. This finding suggests that the same sentiment analysis study could arrive at completely opposite conclusions simply based on the analytical tool used, rather than based on the data.

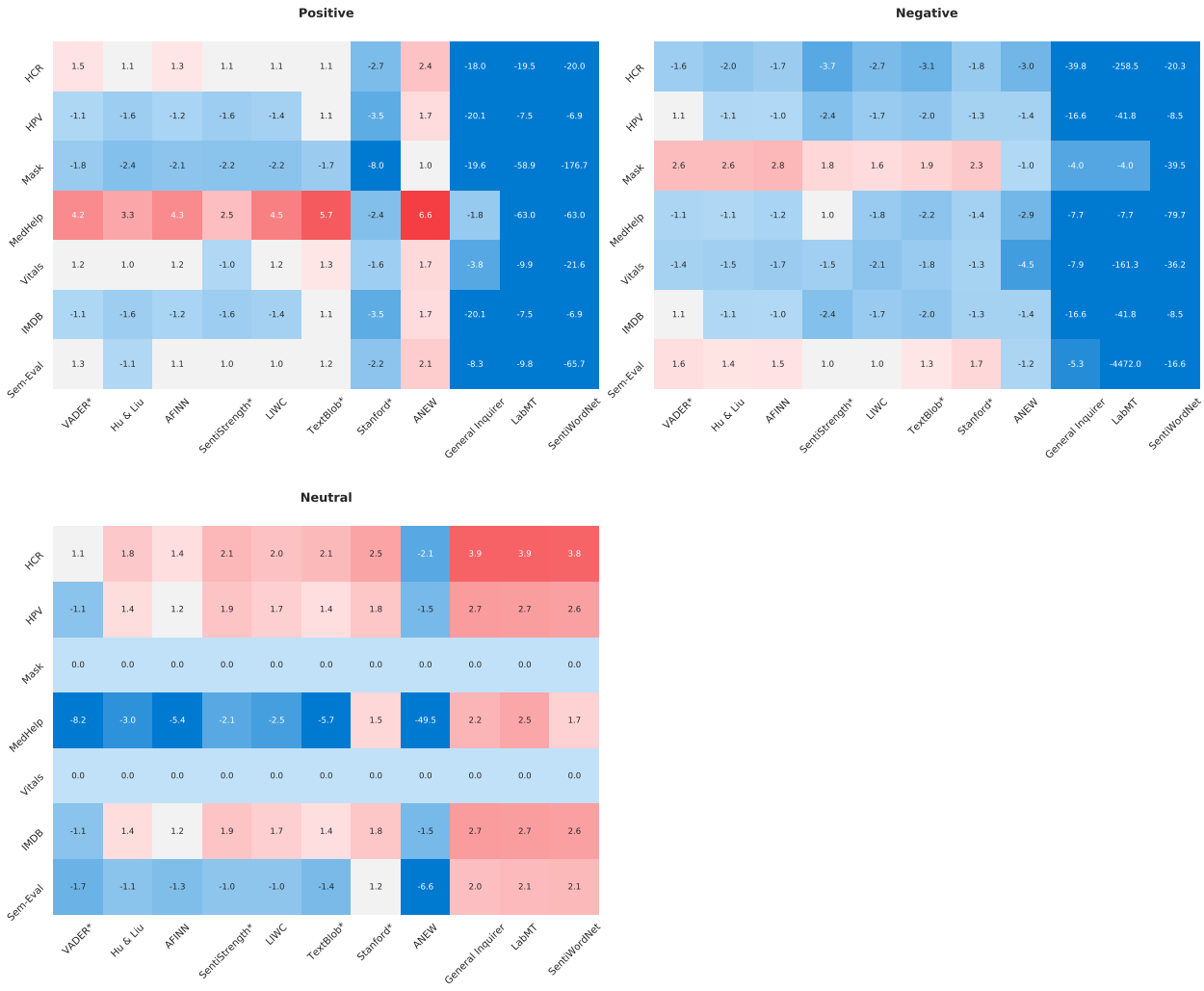

Figure 2.4: Overestimation and underestimation of the tools on the datasets.

Inter-rater reliability

Figure 2.5 reports Fleiss' Kappa scores measuring the inter-rater reliability among the 11 sentiment analysis tools. As shown in Figure 2.5, across the board, the agreements among these tools are very poor, with an average Fleiss' Kappa score of 0.066. The worst scores are observed for the Mask, MedHelp, and HCR datasets, all with a Fleiss' Kappa score lower than 0.04.

	Tweets	Non-tweets
Sentiment analyzers	0.4373	0.6137
Sentiment lexicons	0.3485	0.3734

Table 2.7: Average weighted-F1 score by tool types and data types.

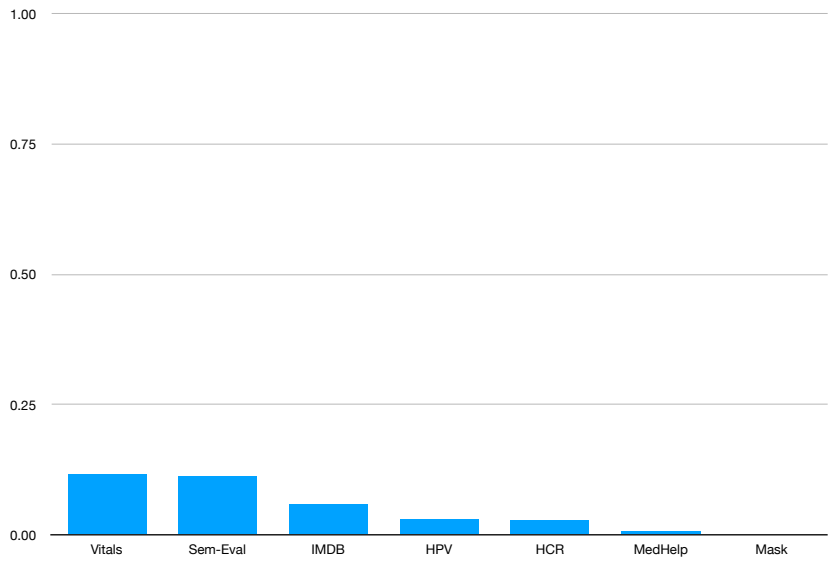

Figure 2.5: Fleiss' Kappa scores on the seven datasets.

Sentiment lexicon coverage

Figure 2.6 shows the characteristics of the six sentiment lexicons (SentiWordNet is not included because it only provides numeric values for positive, negative, and neutral categories, and assumes that a word can simultaneously express different sentiments). Among them, LabMT has the largest vocabulary size (10,222), which is 10 times larger than that of LIWC. General Inquirer and LIWC have roughly equal percentages of positive (45) vs. negative (55) words. On the other hand, AFINN contains fewer positive words than negative words (35.5 vs. 64.5); and Hu & Liu has substantially more negative words (4,782) than postive words (2,006). Only ANEW and LabMT contain a neutral category. ANEW has a very small proportion of neutral words (0.53), though; in contrast, more than 50 of LabMT's lexicon are neutral; only 10 are negative.

Next, we conducted a drill-down analysis to assess the extent to which these sentiment lexicons overlap with one another. Figure 2.7 shows the results. Overall, most lexicons have less than 50% overlapping. For example, only 16 of the words included in Hu & Liu can be found in ANEW, and only 40 of the words included in LabMT can be found in LIWC. While the overlapping rates are low, when a word appears in more than one lexicon, its positivity/negativity/neutrality classification is highly consistent. We only found 32 words that are classified differently. For example, "envy," "poised," "envy," "defeated," and "stunned" are classified as positive words in Hu & Liu while negative in AFINN; and "haunting" and "obsessed" are positive in AFINN while negative in LabMT.

Figure 2.6: Lexicon composition.

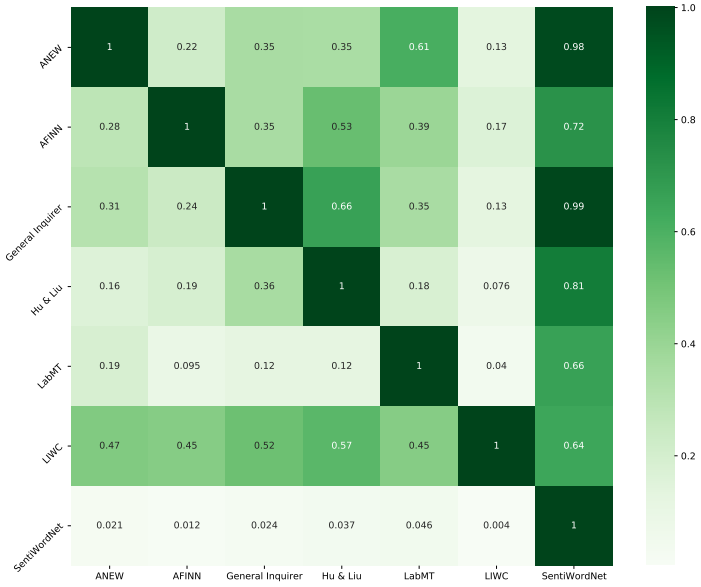

Figure 2.7: Overlapping among sentiment lexicons.

Qualitative error analysis

Among the health-related social media posts, 1,447 were consistently misclassified by all 11 sentiment analysis tools. Through a qualitative error analysis of these misclassified data, we identified two major causes of misclassification: (1) correct sentiment but on wrong subject(s) and (2) inexplicit/indirect sentiment expressions. For the first type, sentimental words (e.g., "excellent," "terrible") appeared in the data but the subject which they were used for was not directly relevant to the topic of the research study. For example, in the tweet, "*Please share this fantastic blog on the facts and myths around the #HPV vaccine. Thanks*", the word "fantastic" was used to describe a relevant blog, which did not provide direct information on the person's opinion regarding HPV. However, most sentiment analysis tools would classify this tweet as positive because of the appearance of the positive word "fantastic." A related issue is the unit of analysis. In the published studies in the literature, sentiment analysis was almost always applied at the post level, e.g., a tweet, a forum message, or a physician review [92]. However, the user may switch between subjects within a single post and express different opinions toward different subjects. For example, "[Physician name] *is Excellent as my ratings above fully described. The only issue I have is the awful receptionist that works at night.*" In this physician review from Vitals.com, the patient expressed opposite opinions toward the physician versus the receptionist. All the 11 sentiment tools that we evaluated, however, do not distinguish such topic turns and thus may not produce accurate sentiment classifications for the specific subject studied. The second type of misclassification originates from inexplicit or indirect sentiment expressions, which are extremely common in social media exchanges. First, sentiments or opinions are often expressed using sarcasm (e.g., "*26 year olds are considered 'children' in Obama's America #tcot #hcr #gopcodered;" analogy*", "*Why would you make 2 years of car payments before actually getting the car?*

THAT is what Obama's #HCR does"), or inexplicit statements that require inference (e.g., "The HPV vaccine isn't just for girls. Here's why: http://t.co/ktOIh ljNMW"). Such subtle sentiment expressions are very difficult to properly interpret for computational algorithms. Second, neutral words, when used in a particular way, may convey sentiments or opinions. For example, "Big Pharma" often has negative connotations in discussions related to controversial drugs, marketing tactics, and the pharmaceutical industry's influence on health policies. The word "big" is however neutral in most contexts and thus is not picked up by the sentiment tools as an expression of opinion. Third, certain social media posts were composed in an obscure fashion, the sentiment of which cannot be readily ascertained without relevant domain knowledge or supple- mental background information. For example, the sentiment expressed in the following tweet, "*Breaking: No deem and pass. Separate vote on Senate bill. Good move. #hcr #healthreform*," cannot be determined without knowing what the Senate bill was and what the implications of separate voting were, yet all sentiment analysis tools would classify it as positive because of the appearance of the positive word "Good." Fourth, a hashtag is a popular means for expressing sentiments on social media, e.g., "*Why wait for the House? Cast your vote on Obama's healthcare bill now; See the trend http://bit.ly/8YtetO #tcot #sgp #dc @#hcr #p2 #politicsm.*" In this tweet, while the sentiment was not explicitly expressed, the hashtag #tcot ("top conservatives on Twitter") gives out that the user was against the Affordable Care Act.

2.2.5 Discussion

Sentiment analysis is a popular method used in informatics research to understand emotions and opinions expressed in social media exchanges. However, most prior sentiment analysis studies used off-the-shelf tools without proper validation for the specific study context [93]. It remains elusive whether these tools can generate accurate sentiment classifications and whether the idiosyncrasies of the tool selected may have a strong influence on study findings

and conclusions. In this paper, we report a comprehensive evaluation of 11 sentiment analysis tools applied to seven datasets. To the best of our knowledge, this is the first study to systematically assess the performance of commonly used sentiment analysis tools in the context of analyzing health-related social media data.

The results show that the existing sentiment analysis tools all performed poorly, with an average weighted F1 score below 0.6. The sentiment classifications generated by more than half of the tools are even worse than what could have been achieved by simply tossing a coin (e.g., SentiWordNet, which has an average weighted F1 score of 0.14). Further, each off-the-shelf tool tends to favor a particular sentiment polarity (polarity bias). For example, SentiWordNet underestimated the positive rates of all datasets evaluated, and General Inquirer consistently overestimated the neutral rates. It is highly questionable whether this level of performance can help researchers accurately classify sentiments in their data to derive useful insights to inform practice (e.g., public health programs, policy making). Further, the machine learning-based analyzers (i.e., Stanford NLP and SentiStrength) did not perform any better than the tools using conventional methods. In fact, the top three performers among the tools that we evaluated are either rule-based (VADER) or lexicon-based (Hu & Liu, AFINN). Furthermore, all tools performed equally poorly when applied to health-related datasets in comparison to generic datasets.

The magnitude of the disagreement between the sentiment analysis tools, as quantified by the extremely low Fleiss' Kappa scores, is also concerning. For example, LabMT underestimated the positivity of the MedHelp dataset by over 60 folds, while TextBlob overestimated it by more than five folds. This means that the very same study, on topics such as whether the public possesses a positive attitude toward mask or vaccination mandates during the COVID-19 pandemic, can reach entirely opposite conclusions simply due to the polarity bias of the sentiment analysis tool used.

Further, the performance of the sentiment analysis tools evaluated is highly variable depend-

ing on the dataset analyzed. Most tools performed better on the Vitals dataset, decently on the Mask dataset, but poorly on the other three health datasets (HPV, HCR, and MedHelp). For example, VADER achieved a weighted F1 score of 0.8 on Vitals, while only 0.41 on HCR. This poor domain transferability suggests that a sentiment analysis tool developed and validated in a particular study context may not be able to achieve the same level of performance when applied in other contexts. Thus, researchers should not blindly trust the validity of a tool simply because it has been used in the past.

The findings of our qualitative error analysis provide useful insights into common causes of sentiment misclassification and how to improve the performance of computational sentiment analysis tools in the future. First, the existing tools use all sentimental words appearing in the text to produce an overall sentiment assessment; none take into account topic turns to determine the specific topic to which a particular sentimental word is applied. This can be very problematic with long texts (e.g., physician reviews from Vitals.com) that commonly contain multiple topics (e.g., quality of treatment, cost, manners of the physician, amenities) with distinct sentiments. Therefore, future sentiment analysis tools need to incorporate topic detection algorithms to ensure that the sentiments classified are truly pertinent to the topic(s) being studied. Second, implicit or indirect expressions such as sarcasm, analogies, and rhetorical questions are, to some degree, a defining characteristic of social media exchanges. To properly interpret such expressions (e.g., the "Big Pharma" example previously discussed), it would require relevant domain knowledge and/or contextual information to be built into the lexicons, rules, or machine-learning models. Lastly, the existing sentiment analysis tools are often inadequate to decode the sentiments expressed through hashtags (e.g., #killthebill), and very few studies to date have included emojis and other popular internet ideograms such as :-) in their analyses [93]. Future sentiment analysis tools may want to incorporate special methods for extracting such information from the data.

Based on our findings, we provide several suggestions for health informatics researchers who

wish to apply off-the-shelf sentiment analysis tools to health-related social media data. First, we suggest that they conduct pre-study evaluation of several tools and choose the one(s) with satisfying performance on their study data. Second, careful error analysis is needed and should be reported in papers in case of systematic errors and bias in results brought by the sentiment analysis tools. Third, we suggest researchers follow a more standardized study design and reporting procedure (e.g., Protocol of Analysis of senTiment in Health [93]). Fourth, we note that even with careful validation, the tools may all produce unsatisfying results as demonstrated in our evaluation. Lastly, we recommend that whenever possible, researchers should prioritize developing customized rules or machine-learning models based on their specific study data, instead of solely relying on off-the-shelf sentiment analysis tools. It is also imperative to develop sentiment analysis tools tailored for health-related social media data.

Our study has several limitations. First, we were only able to include a small number of social media datasets in the analysis that is either publicly available or were provided to us by their developers, which may limit the generalizability of our findings. Most previously published sentiment analysis studies unfortunately do not make their datasets easily accessible, and our attempt to directly request the data from the authors also turned out to be unfruitful. Second, our study only included the most popular sentiment analysis tools that can be readily obtained (either free or for a fee). Their performance thus may not reflect that of all sentiment analysis tools ever developed or under development.

2.2.6 Conclusion

While the suboptimal performance of the off-the-shelf sentiment analysis tools has been previously noted, no research to date has comprehensively investigated their sentiment classification accuracy and common reasons for misclassification. In this paper, we report a

study that comparatively evaluated the performance of 11 widely used sentiment analysis tools on seven social media datasets. The results suggest that none of the tools produced satisfactory results, and their performance varied to a great extent depending on the nature of the dataset. Further, the agreement between these tools is extremely poor, suggesting that the findings of a study can be entirely determined based on the tool selected, rather than based on the data analyzed. Future sentiment analysis research should therefore conduct careful validation of different competing tools for the study context and be aware of the potentially high misclassification rates while drawing their study conclusions. Future methodological development work is also critically needed to improve the performance of computational sentiment analysis tools.

Chapter 3

Study 2: Developing computer-assisted qualitative analysis to understand public and patient concerns toward health-related issues

3.1 Study 2A: What do patients care about? Mining fine-grained patient concerns from online physician reviews through computer-assisted multi-level qualitative analysis

3.1.1 Study summary

Choosing a healthcare provider is a challenging task for patients and caregivers as part of the work they need to take on when managing their health conditions. They often need to consult multiple sources such as other patients, online information, and social circles before making a decision. Online physician review (OPR) websites have been increasingly used by healthcare consumers to make informed decisions in selecting healthcare providers. However, consumer-generated online reviews are often unstructured and contain plural topics with varying degrees of granularity, making it challenging to analyze using conventional topic modeling techniques. In this paper, we designed a novel natural language processing pipeline incorporating qualitative coding and supervised and unsupervised machine learning. Using this method, we were able to identify not only coarse-grained topics (e.g., relationship, clinic management), but also fine-grained details such as diagnosis, timing and access, and financial concerns. We discuss how healthcare providers could improve their ratings based on consumer feedback. We also reflect on the inherent challenges of analyzing user-generated online data, and how our novel pipeline may inform future work on mining consumer-generated online data.

3.1.2 Introduction

Choosing the right healthcare provider has been a challenge to many patients due to inherent information asymmetry between the two parties. As a result, patients often seek advice from friends and family who had similar conditions and experiences [89, 99]. This pressing information need has given rise to online physician rating (OPR) websites, where millions of patients can share experiences by reviewing and evaluating their physicians. It is estimated that popular OPR websites, such as Vitals.com and RateMDs.com, are consulted by at least 30 Internet users in the U.S. and have significance influence on people's choices of healthcare

providers [47].

Data from OPR websites (henceforth called OPR data) cover a variety of inforamtion. This includes physician profiles (specialty, experience, accepted insurance, etc.), overall satisfaction ratings (1-5 stars), break-down ratings (along multiple dimensions such as competence, wait time, bedside manner, etc.), and open-ended reviews written by patients. This data provides a unique lens through which many stakeholders can obtain insights. For example, healthcare providers can better understand patient concerns to improve quality of care; health informatics researchers can gain better understanding of consumers' information needs; healthcare consumers such as patients and caregivers can be empowered through better information access; government agencies can design more comprehensive healthcare quality assessment surveys [1].

OPR has been increasingly studied in the research community. Early studies focused on analyzing consumer ratings as these structured data can be easily processed at scale. One type of work cross-checked consumer ratings against professional surveys and clinical performance, and they discovered inconsistent results. Gao et al. found that ratings on RateMDs and measurements from the official state medical board had significant positive correlation with an increasing support from 2005 to 2010 [82]. However, Daskivich et al. found that online ratings failed to correlate with experience than objective treatment outcomes, and therefore consumer ratings may reflect different aspects of concerns than those in official surveys. Such inconsistency is likely due to the fact that OPRs are based more on consumers' subjective experience than objective treatment outcomes, and therefore consumer ratings may reflect different aspects of concerns than those in official surveys [59].

Compared to ratings, free-text reviews in OPR websites are more nuanced and carry richer information about patient concerns. However, the sheer amount of unstructured reviews makes it infeasible to conduct exhaustive analysis. As a compromise, previous researchers take one of two approaches. The first approach samples a relatively small set of reviews from

the big OPR data for a focused qualitative analysis. For example, Lopez et al. analyzed 712 reviews from Yelp and RateMD and identified three major themes: technical competence, interpersonal manner, and system issues [129]. Kilaru et al. used a grounded theory approach to analyze 1,736 reviews of emergency department (ED) care on Yelp and found that similar topics are shared between Yelp reviews and those in official surveys [116]. These studies, while providing deep insights into patient concerns, only covered a small sample of all reviews. To scale up the analysis, the second approach employs machine learning techniques such as statistical topic modeling to extract topics (each topic consisting of a list of keywords) from large-scale consumer reviews. For example, Wallace et al. adopted the three themes identified in Lopez et al. and applied topic modeling on nearly 60,000 reviews from RateMD [199]. A recent analysis discovered three general topics (hospital-level services, communication skills, and professional skills) from a Chinese OPR website [152]. While these studies demonstrate the potential of computer-assisted qualitative analysis [53], the extracted topics were often coarse-grained and provided only the high-level categories of topics without identifying any detailed aspects under each top topic. Indeed, interpreting topics extracted from consumer-generated reviews can be challenging [49], especially when review texts have short lengths, correlated topics, and nested subtopics [187].

3.1.3 Material and methods

Data description

Vitals is one of the largest OPR websites for healthcare consumers to provide or access evaluations of physicians in the U.S [113]. The site has 127,300 unique daily visits according to Google Trends. The site provides basic information on physicians, such as their locations, gender, and year of experience, etc. Patients are able to score a doctor on a Likert scale of 1 (poor) to 5 (excellent), write a review and selectively make a detailed quality rating across eight dimensions: Wait Time, Easy Appointments, Promptness, Friendly Staff, Accurate

Diagnosis, Bedside Manner, Spends Time with Patients, Appropriate Follow-up. In this study, we collected and analyzed 1,065,631 OPRs posted from January 1, 2008 to November 4, 2018 for 102,540 family physicians in the U.S. on Vitals.

Method pipeline

We employed a multi-level qualitative analysis method pipeline. The basic idea is to take a top-down approach to mining a large-scale review corpus. We first identified coarse-grained, high-level topics, and then identified fine-grained, low-level subtopics (or detailed patient concerns) under each topic. To scale up the analysis to a large corpus, we combined manual coding with machine learning in both stages. For our text analysis, we only included the 1-star and 5-star reviews that have more than 20 words. We chose this subset because they represent the majority of the reviews and are long enough to be informative. In addition, 1-star and 5-star reviews convey direct negative and positive emotions, while the moderate reviews (2, 3 and 4-star) often convey mixed feelings, which is challenging to disentangle.

Mining coarse-grained topics

To identify coarse-grained topics, we conducted qualitative coding on a sample of reviews, and then used supervised machine learning to generalize the codes to all reviews. We did not use topic modeling to automatically discover coarse-grained topics, because algorithms like latent Dirichlet allocation extracted uninterpretable topics with mixed content in pilot experiments. Indeed, these algorithms work well when topics are well separated [187]. However, themes in OPR reviews are often mingled. For example, dissatisfied consumers often simultaneously complain about lack of clinical competence and bad interpersonal manners. (Semi-)supervised topic modeling is not pragmatic either as it assumes that qualitative analysis has been done in the first place [199].

To ensure that each review contains enough information for qualitative coding, we only considered reviews with at least 20 words. This resulted in a corpus with 207,029 free-text reviews.

1. Qualitative coding of reviews: We used concepts from a validated patient complaint taxonomy initially proposed by Reader et al to guide our coding [161]. We chose this taxonomy because it was built through a systematic synthesis of patient complaint literature and has been validated and used in many patient satisfaction studies17. 200 reviews were randomly selected and coded by two annotators separately. The two annotators discussed to resolve disagreements and reached an agreement ratio above 80%. In this annotation stage, the annotators found that the three concepts in the taxonomy (management, clinical, and relationship) captured all the topics in the reviews and no new topics emerged. The two annotators separately coded another 400 reviews, resulting in a set of 600 annotated reviews. In this training set, 59.3% were labeled as including clinical topics, 34.2% management, and 75.5% relationship.

2. Supervised review classification: We used the 600 annotated reviews as training data to train text classifiers that assign topics to unannotated reviews. A review was represented as a feature vector by taking the average of its word vectors, known as a continuous bag-of-words representation [138]. Words were represented as 100-dimensional vectors trained by the word2vec algorithm on the review corpus. We trained one classifier for each topic, so that each classifier decided whether a review belongs to a topic. This allows a review to have multiple topics. We chose gradient boosted decision trees as the underlying classification model, as it showed higher accuracy than support vector machine or random forest. Two hyperparameters, maximum depth of trees and minimum sum of instance weights in a leaf, were optimized for each classifier. Under 10-fold cross validation, the classifier achieved 84% F1-score on management, 86.7% on clinical, and 92.5% on relationship. These machine predictions are remarkably accurate since they are about the same as human agreement rate.

3. Estimating word-topic relatedness. We measured the relatedness between a word and a topic as the probability of a word being classified into a topic, according to the

corresponding topic classifier.

Mining fine-grained concerns

To identify fine-grained concerns (or aspects) under each topic, we ran a clustering algorithm on topic-related words, and then examined and annotated these word clusters. Here we adopted word clustering instead of manual coding as we found empirically that such an algorithm could already discover interpretable aspects. This is likely because latent aspects are almost uncorrelated under the same topic (i.e., conditionally independent [61]) and give rise to distinct word clusters.For each topic, we clustered 3,000 words (10% vocabulary size) with the highest word-topic relatedness computed in (3).

1. Unsupervised word clustering: Given topic-related words under each topic, we applied k-means algorithm over the word vectors (learned in Step 1). Euclidean distance between two vectors is used as the distance measure. These clusters represented candidate aspects under each topic that expressed fine-grained patient concerns. To avoid omitting aspects, we set k = 20 clusters for each topic, which is more than twice the number of aspects in previous work [161].

2. Qualitative coding of word clusters: Two annotators independently examined 10 words closest to each cluster centroid to determine its meaning. Inspired by the divide-and-merge methodology for clustering [51], we manually merged clusters with similar meaning. If two clusters exhibited opposite attitudes towards the same subject matter, they were also merged. Our manual coding was also guided by Reader et al.'s taxonomy [161].

3. Estimating review-aspect relatedness: We measured the review-aspect relatedness as the reciprocal of cosine distance between the review document vector to the cluster centroid of aspect. Since a review may talk about more than one aspect, we calculated the review-aspect distance for all aspects under that topic and assigned the normalized

relatedness to a review instead of assigning the closest aspect to it.

The overall method pipeline is depicted in Figure 3.1.

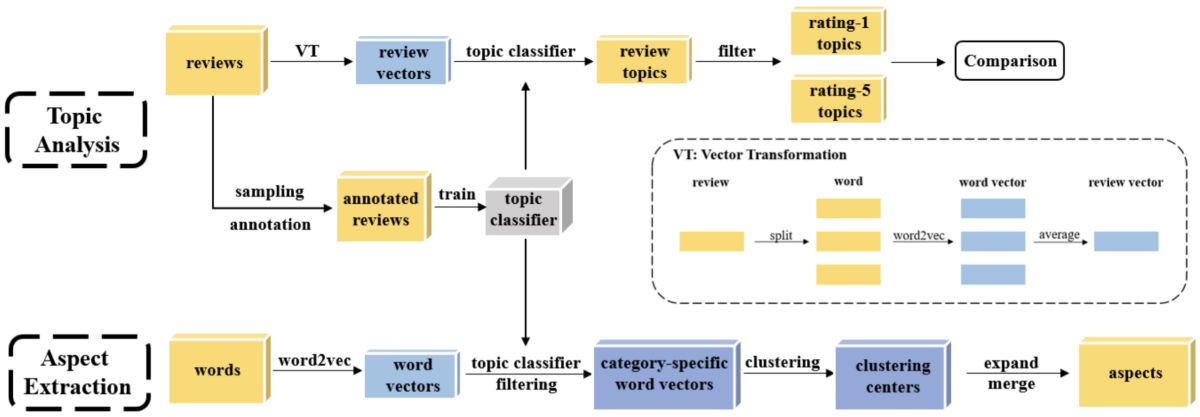

Figure 3.1: Method pipeline.

3.1.4 Results

This section will first provide an overview of the dataset through consumer rating analysis, and then report the coarse-grained topics and fine-grained aspects identified through our novel computer assisted qualitative coding process.

Consumer ratings

The rating distribution at the review-level is J-shaped, with 66% being 5-star, 16% 1-star and the rest 18% in the middle, and is consistent with previous findings [113]. At the physician level, the average rating is 4.039 and the standard deviation is 0.926, indicating that physicians tend to receive favorable ratings overall. The average number of reviews a

physician received is 10.44 and the standard deviation is 13.2. 24.5% of the physicians only received one or two reviews, suggesting a highly skewed distribution of the number of reviews at the physician level.

While the website allows users to rate physicians on 8 sub-categories listed above, more than half of the reviews did not have any of the 8 categories rated. Among them, wait time and follow-up have higher unfilled proportions. Moderate reviews (2, 3&4) tend to have more unrated subcategories compared to extreme reviews (1&5). Specifically, 41% of the 1-star reviews have all of the 8 categories unrated, 45% for 5-star reviews, while 72% of the 3-star reviews have all of the 8 categories unrated and 65% for the 4-star reviews.

Coarse-grained topic analysis

The reviews were classified using the machine learning model to decide whether they include the three topics: relationship, clinical, and management. **Relationship** refers to interaction between patients and physicians. This could include their communication and physicians' empathy toward patients. **Clinical** refers to patients' perceived quality of care. **Management** refers to institutional managerial issues. For example, patients complained about long waiting time and difficulty scheduling appointments.

Among 207,029 1-star and 5-star reviews with at least 20 words, 193,360 (93.4%) were predicted relevant to relationship, 146,358 (70.7%) to clinical, while only 78,391 (37.9%) were predicted relevant to management. This suggests that overall health consumers wrote more about physician-patient relationships and clinical issues than management when evaluating physicians online. Nearly one fifth of the reviews (43,331, 20.9%) were classified to include all the three topics. 126,103 (60.9%) reviews talked about 2 topics, 35,909 (17.3%) mentioned 1 topic, and 1,686 (0.8%) did not belong to any topic. Those reviews that do not include any of the three topics mostly provide general evaluations such as "*He is over all a very good dr. I have been going to him for over 20 years. I have no complaint.*"

Topic	Words	Example	Proportion
Relationship	listening, attentive, respectful, receptive, interrupt, hurry, rush, belittling, empathetic, unconcerned	Dr. X is one of the nicest Dr's I've met here. He took the time to listen completely without interruption and he explained in a way and could understand.	93.4%
Clinical	anemia, dangerously, remedy, beneficial, diagnoses, anti-inflammatory, insightful, gallbladder, evaluation, recommendations	I was initially upset because he wanted to do a lot of workup on my heartburn, but I am glad he did. It turns out it was my heart and not acid reflux. Thank you!	70.7%
Management	rescheduled, 8am, appointments, follow-ups, billing, insurance, understaffed, chaotic, expired, wednesday	Once a patient it's becomes increasingly hard to get an appointment or seen in between the "follow-up" visits. It's all about the dollar.	37.9%

Table 3.1: Coarse-grained topics.

Table 3.1 presents the three topics, words highly related to the topics, selected examples and the proportions. The words have high correlation with the corresponding topics are selected based on word-topic relatedness in Method (3). We replaced real physician names with X to preserve privacy. We kept the misspellings, grammatical errors and capitalization as they appeared in the original dataset. To find the trigger of leaving positive/negative reviews, we made a comparison on the topic distribution of 5-star and 1-star reviews as shown in 3.2. Both clinical and relationship related issues appeared slightly more in 5-star reviews than in 1-star reviews. However, management was discussed much less in 5-star reviews as compared to in 1-star reviews. Only around 20% of the 5-star reviews discussed management, while more than 60% of the 1-star reviews discussed management. The proportion of management related reviews in 1-star reviews and 5-star reviews is significantly different ($p < 0.05$).

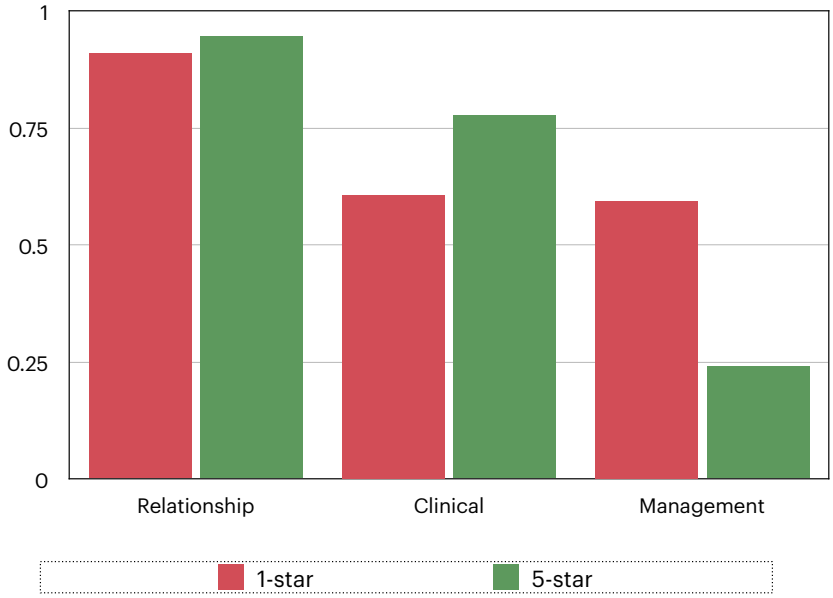

Figure 3.2: Topic distribution in different ratings.

Fine-grained aspect analysis

To extract the fine-grained aspects under each topic, we combined unsupervised word vector clustering and qualitative coding. We summarized our findings in Table 3.2–3.4. Since a review can include multiple topics and aspects, the review examples we put under one aspect can also be under several other aspects. Note that the sum of aspect proportion under a topic equals 1 because we assigned the normalized relatedness of each aspect to a review.

There are four aspects identified under relationship: Patience, Communication, Respect and Compassion as shown in Table 3.2. **Patience** refers to whether physicians spend time with

Aspect	Keywords	Examples	Proportion
Patience	hurry, rush, examines,forgets, interrupting, cuts, interrupts, intently, dismisses, patiently	I have the up most respect for Dr. X. She iskind, patient & her appointments are prompt. She answers all your questions & is not hurried. I believe she schedules patients 30 minutes apart. 1 visit with her is like multiple visits with an Urgent care doctor.	30.3%
Communication	listening, addressing,dismiss, evaluate, brushed, voiced, hears, brush, receptive, express	I really felt he had an excellent presence and extremely helpful. He took time to listen to my concerns and cared about my issues. I would highly recommend him to family/friends.	25.0%
Respect	belittling, patronizing,sarcastic, smug, abrasive, unconcerned, unsympathetic, hostile, combative, argumentative	Very arrogant and patronizing, also quite inappropriate and rude at times. Did not care to look for a resolution to my ailment. After two years of this my last interaction with him made me switch physicians.	23.1%
Compassion	empathetic, thoughtful, respectful, approachable, considerate, insightful, informative, personable, conscientious, sympathetic	Dr X truly makes you feel you are his only patient..He is empathetic sympathetic and very kind...Many days we have cried together...God could not created a better human being to be a Dr to administer care for the sick...I am so grateful to be a patient	21.9%

Table 3.2: Aspects under Relationship.

patients in person. Some patients felt being rushed during clinical encounters and were often ignored or interrupted. **Communication** refers to the quality of patient-provider conversation. Patients commented on whether physicians listened to and addressed their questions. **Respect** refers to whether patients were treated in a respectful manner. For instance, some patients reported that their physicians were arrogant and abrasive. **Compassion** refers to the tenderness, compassion and sympathy toward patients. For example, some patients described their physicians as empathetic and sympathetic.

Our cluster analysis shows that when patients talk about relationship-related aspects, they tend to write using more emotional terms and strong adjectives to express their dissatisfaction or compliment, such as "*He is empathetic, sympathetic and very kind" in the example for Compassion and "Very arrogant and patronizing, also quite inappropriate and rude at times*" in the example for Respect. In addition, though we manually merged the clusters, some of the relationship-related aspects are not exclusive from each other. For instance, in the example for Communication, "*He took time to listen to my concerns and cared about my issues*" also reflects the patience and compassion of the doctor. Besides, we also found when

talking about communication issues, patients are more likely to mention whether physicians listen to their concerns instead of whether the doctors express precisely, which echoes the importance of listening in doctor-patient communication as previous work suggested [109].

For the topic clinical, five aspects were identified: Treatment, Diagnose, Medication, Personal Conditions and Professional skills, as shown in Table 3.3. **Treatment** refers to how physicians treat patients' diseases. For instance, patients described the kinds of treatment plans and whether they turned out to be effective. **Diagnose** refers to the assessment and judgements of clinical symptoms. For example, patients described how the physicians diagnosed them and whether they have been misdiagnosed. **Medication** refers to the prescription and administration of medications. Patients listed the names or types of medications that they were prescribed such as anti-depressant and anti- inflammatory. **Personal conditions** refer to patients' personal health conditions, medical history and symptoms. **Professional skills** refer to physicians' overall clinical competence. Patients generally used adjectives to describe their perceptions of the clinical competence of physicians. For example, they may describe a physician as "meticulous", "well-informed" or "astute". We observed that in clinical-related OPRs, the five aspects tend to be discussed collectively. For example, the following review, "*39 year old male here. I have been dealing with occasional hip pain on and off for years. Dr. X did a physical exam and X-rays. I was diagnosed with bursitis and tendinitis. Some anti inflammatory meds were prescribed which worked. This was good news since I really didn't want to pay for an mri or have surgery. I realize that not everyone may not be so lucky with their diagnosis. He spent a lot of time with me and yet I still feel like I was in and out. His staff was kind and courteous. I rarely write reviews but my experience was just too good to not mention*", first describes the whole procedure from providing personal medical history (occasional hip pain), being diagnosed (bursitis and tendinitis), and to being prescribed medications (anti-inflammatory). At the end, the review makes an evaluation of the doctor's overall professional skills based on the previous procedures.

Aspect	Keywords	Examples	Proportion
Treatment	possibilities, protocol, appropriately, prognosis, method, symptoms, remedy, pharmaceuticals, determining, effectively	Dr. X diagnosed and effectively treated a very burdensome problem that many previous physicians could not help me with	23.8%
Diagnose	diagnoses, conclusions, direction, prognosis, recommendations, findings, possibilities, assessments, evaluation, judgements	Dr X did not listen to our needs. She was very full of herself. misdiagnosed sinus infection as a virus. Had to go to another doctor to get treated.	22.2%
Medication	anti-inflammatory, prednisone, inflammatory, anti-depressant, zoloft, depressants, toxic, temporary, topical, statins	I went to her throughout my pregnancy. She recommended antidepressants such as Zoloft which cause birth defects. She had no idea of what she was doing. Even the nurses that worked with her told me that I should switch doctors.	19.2%
Personal conditions	pulmonary, ovarian, gallbladder, colon, cancerous, artery, lymph, cervical, fluid, blockage	Definetely i do not recommend this dr. to nobody, I had my gallbladder removed last year and this surgery went bad. I had unexpected life threatening complications. She never took the time to figure out what she did wrong in the surgery. Result of this procedure i was admitted to hospital 5 and a half months . weeks of being intubated. i also have permanently health impairments.	17.6%
Professional skills	diligent, insightful, intuitive, keen, astute, meticulous, forthright, realistic, well-informed, precise	Dr. X is thorough, insightful, kind and accurate. He quickly diagnosed my case and proposed a plan and solutions. I wish he were available as a primary care doctor–he is a top flight emergency physician!	17.2%

Table 3.3: Aspects under Clinical.

We identified five aspects that fall under management: Timing and access, Bureaucracy, Finance and billing, Service issues and Staff and resources, as shown in Table 3.4. **Timing and access** refer to timely and easy access to healthcare services. For example, patients commented on their waiting time to be seen by doctors, and ease of scheduling and rescheduling appointments. **Bureaucracy** refers to the administrative rules with the healthcare organization. For instance, it may involve having a prescription verified and getting a signature or authorization from the office. **Finance and billing** refer to the financial components of healthcare services such as insurance, billing and payment. For example, users shared their experience of being overcharged or having difficulty in their billing processes. **Service issues** refer to hospital services provided for patients in their encounters. These include follow-ups and resolving issues. For example, a patient wrote that the billing code was entered incorrectly, and no one has followed up and resolved this problem. **Staff and resources** refer to whether the healthcare organization has adequate and well-trained staff and appropriate

Aspect	Keywords	Examples	Proportion
Timing and access	noon, wednesday, tuesday, thursday, rescheduled, 8am, app, 10:30, notified, reminder	Once a patient it's becomes increasingly hard to get an appointment or seen in between the "follow-up" visits	22.1%
Bureaucracy	verified, expired, insist, declined, processed, application, signature, issued, authorizations, approve	can't get anyone to ever call me back for follow up and for help with getting prescriptions sent out or verified.	21.9%
Finance and billing	charging, co-pays, cards,owed, 250, fees, payments, refund, agency, deductible	HE IS EXCELLECT, JUST VERY UNAWARE THAT HIS STAFF IS CHARGING FULL ENGORGED OFFICE PRICES FOR CASH PAYMENTS, DESPITE INSURE COMPANIES ONLY PAY ABOUT A THIRD AND ITS ACCEPTABLE FOR THE INSURED!!!!	21.2%
Service issues	processes, informs, speed, follow-ups, monitors, consultations, adjusts, receptive, conflicting, resolution	Dr. X is a caring and problem solving doc. she always support and provides her best consultations at par. she and her nurse practitioner provides support even if we had left them a message and they phoned us back providing the refer and consultations.	17.5%
Staff and resources	inefficient, unwelcoming, sloppy, untrained, understaffed, uncooperative, inattentive, clerical, chaotic, staff'	Dr. X is professional, engaging and pleasant. The receptionists and other low-level staff are, however, quite unprofessional. They all need training on how they handle people and how to conduct themselves in an office or she will lose patients based purely on her staff's behavior	17.3%

Table 3.4: Aspects under Management.

resources. Among the five aspects, timing and access, bureaucracy and finance and billing are mentioned most. We also noticed that management-related OPRs are significantly longer than relationship-related OPRs and clinical-related OPRs, which could be attributed to a more detailed description when talking about aspects under management.

Figure 3.3 showed the distribution of different aspects mentioned in 1-star and 5-star reviews. Each topic-relevant review was assigned one aspect. Overall, a physician's patience, compassion, professional skills, accurate diagnosis, effective treatment and good services are appreciated by patients in positive (5-star) reviews. In negative (1-star) reviews, patients often refer to their personal conditions and medication to contextualize their complaints, especially on lack of respect and bureaucratic processes.

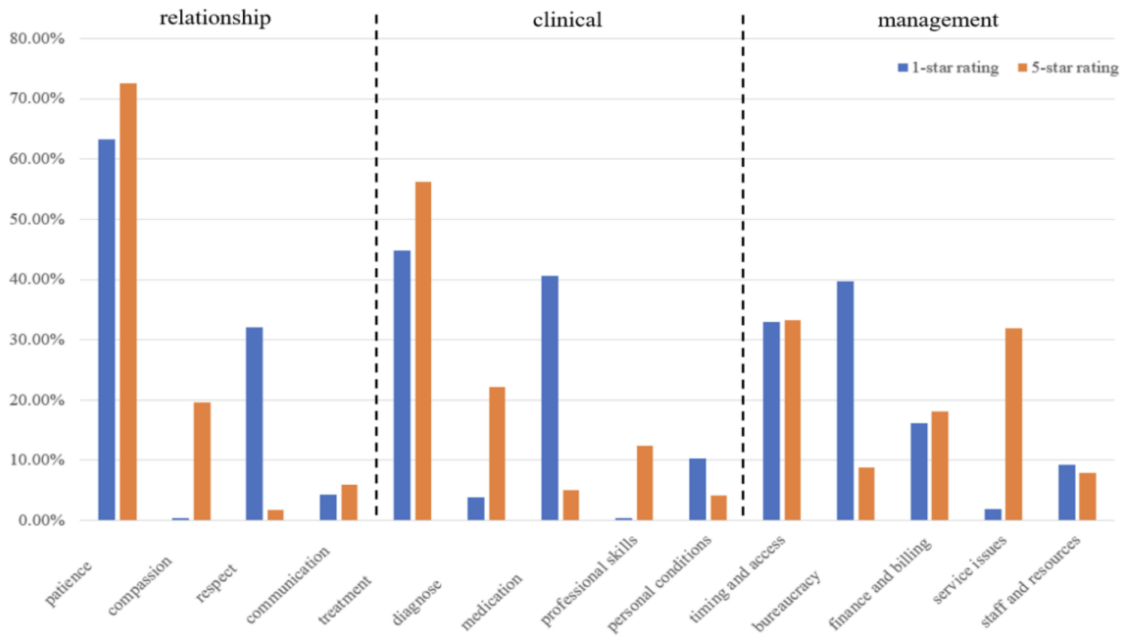

Figure 3.3: Aspect distribution in different topics and ratings.

3.1.5 Discussion

In this paper, we developed a novel computer-assisted qualitative coding methodology to mine coarse-grained topics and fine-grained aspects from consumer-generated OPR data. Through manual coding and supervised machine learning, we extracted three major topics from OPRs: management, clinical, and relationship. Through unsupervised word vector clustering and qualitative coding, we further identified fine-grained aspects such as timing and access, diagnosis, and communication.

A general methodology for fine-grained analysis of consumer-generated texts. Free-text patient reviews are often mixtures of factual topics intertwined with personal feelings across multiple dimensions and granularities. To fully uncover the fine-grained semantics from texts, it is unrealistic to solely rely on unsupervised algorithms such as topic modeling

or their semi-supervised variants that only take one round of human input. Instead, an interleaving of human coding and machine learning is essential to achieve nuanced understanding of these texts. This work introduces a novel analysis methodology that takes a divide-and-conquer approach: it first divides the content into coarse-grained topics, and then zooms in on each topic to locate fine-grained concerns. Human coding is amplified through supervised learning in the first stage and aided by unsupervised learning in the second stage. Together, the methodology effectively interleaves a small but essential amount of human effort with the large-scale processing capability of machine learning in a qualitative analysis task. This general methodology can be useful in a variety of scenarios where fine-grained analysis of consumer-generated texts is needed.

Implications for healthcare service quality improvement.

At the coarse-grained topic level, we found that relationship was discussed in 93.4% of the reviews, suggesting that patient-provider relationship is of high-priority for patients. In addition, we found that users discussed management-related topics much more often in 1-star reviews than in 5-star reviews. A hypothesis to explain this phenomenon is that poor management would greatly affect patients' experience with healthcare service, while good management is less noticeable and thus not frequently mentioned in favorable reviews. This finding echoes with previous work which suggests that "[...] *80-94 percent of the damage done by poor service quality is traceable to managerial actions or the system set up by management.*" [77] Therefore, though management is not directly related to clinical performance, it could be the triggers for healthcare consumers to leave unfavorable reviews online. These findings also suggest that the inconsistency between online physician ratings and objective clinical performance could be in part due to the fact that they are evaluating very different aspects. Healthcare providers and government agencies should consider better ways of measuring healthcare consumers' satisfaction with their services by gaining insights from consumer-generated online data and including more non-clinical related aspects. Through unsupervised word vector clustering and manual coding, we were able to identify fine-grained

aspects that greatly complement OPR literature by providing a granular and richer description of healthcare consumers' narratives on OPR websites, which shed light on more substantial solutions to improve healthcare service quality. We found that consumer-generated OPR data encompass a wide range of healthcare service aspects, including timing and access, finance and billing, diagnoses, medication, and communication, etc. In management related reviews, timing and access, bureaucracy and finance and billing were mentioned more often than staff and resources and service issues. This indicates that healthcare consumers discussed more about whether they had timely and easy access to healthcare services and whether their interaction with the healthcare organization was smooth.

Limitations and future work

First, we only studied one OPR website and the findings may not generalize to other OPR websites with different designs or target users. Second, we only included family physicians in this study. Patients may value different aspects of family physicians compared to other specialists such as surgeons and dentists. We plan to conduct cross-platform and cross-specialty comparisons in our future work.

3.1.6 Conclusion

We developed a novel computer-assisted qualitative coding method to mine multi-level patient concerns from a large-scale heterogeneous OPR corpus. We identified coarse-grained topics (management, clinical, relationship) as well as fine-grained aspects (e.g., bureaucracy, diagnosis, communication) which provide more granular and richer information of patients' evaluation of healthcare quality online. Our results compliment previous OPR research by contributing the multi-level patient concerns and the novel method for mining large-scale heterogeneous consumer-generated texts.

3.2 Study 2B: Why do people oppose mask wearing? A comprehensive analysis of U.S. tweets during the COVID-19 pandemic

3.2.1 Study summary

During the COVID-19 pandemic, facial masks were debated heatedly as a personal protective equipment in the United States. As a result, the mask adoption rate in the United States was also lower than in other countries. The rationales for opposing mask wearing and how public perceptions evolve over time are not well studied to inform more effective public health communication. In this study, we analyzed a total of 771,268 U.S.-based tweets between January to October 2020. We developed machine learning classifiers to identify and categorize relevant tweets, followed by a qualitative content analysis of a subset of the tweets to understand the rationale of those opposed mask wearing. We found that while the majority of the tweets supported mask wearing, the proportion of anti-mask tweets stayed constant at about a 10% level throughout the study period. Common reasons for opposition included physical discomfort and negative effects, lack of effectiveness, and being unnecessary or inappropriate for certain people or under certain circumstances. The opposing tweets were significantly less likely to cite external sources of information such as public health agencies' websites to support the arguments. The results may inform better communication strategies to improve the public perception of wearing masks and, in particular, to specifically address common anti-mask beliefs.

3.2.2 Introduction

The coronavirus disease 2019 (COVID-19) pandemic has caused significant morbidity and mortality across the globe. On December 28, 2020, there were 441 861 new confirmed cases worldwide, with 145,959 of these being in the United States [209, 43]. While scientific research has advanced our knowledge of the disease, new therapeutic treatments have been developed and vaccines are now approved and available, widespread adoption of individual protective behaviors, such as wearing personal protective equipment and practicing social distancing, remains crucial to reducing the spread of COVID-19 [43]. Despite a handful of studies questioning the effectiveness of community mask wearing [36, 145], it has been shown that facial masks, even if homemade, can lead to a decrease in mortality by more than 20% if worn by more than 80% of community members [71].

The benefits of facial masks can only be realized when most people wear them, which is known as universal mask use [185]. While several countries (eg, Singapore, South Korea, China) have achieved this goal [114], promoting widespread mask wearing in the United States has encountered substantial obstacles. Besides cultural norms dictating that only the sick wear masks, the anti-mask opinion held by some government officials, and shifting positions by U.S. and international public health authorities, have added confusion to the debate. In particular, the U.S. Centers for Disease Control and Prevention (CDC) initially recommended against public mask wearing on February 27, 2020 [44]. However, on April 3, the CDC reversed its position to instead recommend universal mask use [41]. Then, on April 6, the World Health Organization (WHO) issued an advisory stating that healthy individuals do not need to wear masks, directly contradicting the CDC's recommendation.10 Further, there has been significant variation across state and county health departments on mask-wearing policies, and the debate on whether or not there should be a national mask mandate in the United States remains unsettled [211].

While surveys by The New York Times, the Pew Research Center, and the CDC reported a relatively high self-reported mask use rate in the United States (59%, 65%, and 74.1%, respectively), the actual adoption rate is questionable. For example, in the same survey conducted by the Pew Research Center, only 44% of the participants reported that members in their communities were actually wearing masks all or most of the time when in public, suggesting that social desirability bias may be affecting self-reported rates [41]. Further, all currently available surveys used simple yes/no questions on mask wearing without soliciting the rationales behind the opposing opinions [114, 41, 46]; only a handful of news outlets and advocacy groups have provided excerpts from the public or conjectured on why some people refused to wear masks [32, 110, 96, 175]. Finally, most of the existing surveys were carried out at sporadic times for cross-sectional analysis in limited geographic areas. Continuous monitoring of public perception across the country is rare, leaving a knowledge vacuum of understanding how the public attitudes toward mask wearing have evolved over time since the beginning of the pandemic [114, 41, 46].

To address these gaps, we analyzed a large Twitter dataset collected in the United States from January to October 2020 to answer the following research questions (RQs):

1. (a) What is the general public's attitude toward mask wearing in the United States?
   (b) How has the general public's attitude changed over time as the pandemic progressed?
2. Among those expressing an anti-mask opinion, what are their concerns or justifications?
3. What is the external source of information shared to support the pro- or anti-mask arguments?

Based on the results of available conventional surveys, we hypothesize that the general public's attitude toward facial masking expressed through tweets would be generally positive,

even though unfavorable viewpoints would not be uncommon. Further, this attitude would have shifted over time as a result of changing CDC guidelines and local mask-wearing policies and how such policies are enforced. We also hypothesize that anti-mask tweets would be less likely to cite external sources of information especially from public health authorities.

3.2.3 Material and methods

We used a computer-aided qualitative analysis approach that combines machine learning and qualitative content analysis. To answer RQ1 (changing attitude), we trained a machine learning classifier to label personal opinions regarding mask wearing and to examine its evolution over time. To answer RQ2 (concerns or justifications for opposition), we conducted an in-depth qualitative content analysis of a random set of anti-mask tweets to examine the common beliefs held by those who opposed mask wearing, as well as the reasoning behind such beliefs. To answer RQ3 (sharing of external information), we quantitatively analyzed the external evidence cited in the tweets, eg, by calculating the proportion from public health authorities such as WHO and the CDC. The overall analytical flow is exhibited in Figure 3.4. To protect the privacy of individuals, we paraphrased all tweets presented in the following sections, instead of directly quoting the original tweets.

Data collection and preprocessing

Retrieving tweets

As this study concerns the public attitudes toward mask wearing in the United States, only those geo-coded tweets falling into the coordinates of {-173.847656, 17.644022, -65.390625, 70.377854}, which approximately represents the continental United States, were included. Geocodes are the main source of information that researchers could use to determine user locations. While user profiles are available through the Twitter API, we did not opt to use

"mask" OR "face cover" OR "cloth cover" OR "face cloth" OR "mouth cover" OR "nose cover" OR "facial cover" OR "nose cloth" OR "eye cloth" OR "mouth cloth"

Table 3.5: Search keywords

them for inferring geolocations because users' locations specified at the point of registration may not match their locations when relevant tweets were posted. Therefore, only the geolocations attached to tweets meeting our criteria were used. Based on the estimate provided by Twitter, approximately 1% to 2% of Twitter users opt to allow their geolocation information to be tracked [14].

The data collection of this study covered a 10-month period between January 1 and November 1, 2020. It was conducted by leveraging the official Twitter API version 1.1, which continuously retrieves all tweets meeting the search criteria from a stream of general tweets that are located in the United States provided by Twitter [195]. The keywords used are detailed in Table 3.5, which were developed based on a manual review of sample tweets in addition to examining the search terms employed in prior research [167].

Further filtering

Social media data retrieved through keywords search may contain a substantial amount of irrelevant content, eg, "face cloth" may refer to facial washcloth for makeup removal [117]. To remove such noise, we manually analyzed a random set of tweets to examine the characteristics of irrelevant posts (a total of 200 tweets were reviewed; saturation was achieved after coding about 70 tweets). The results show that the majority of such tweets fall into the following 2 categories: (1) those related to mask manufacturing or product advertisements or that referred to other meanings of the word mask (eg, "*The government has been masking the fact that the it is a failure*"); and (2) those pertinent to mask wearing but did not express any personal attitude (eg, tweets that merely shared a URL with no personal opinions explicitly stated), or the attitude is difficult to discern from the content of

the tweet (eg, "*Should you wear a mask?? COVID facemask comfortmask. Read this blog: URL*").

To remove such tweets from further analyses, we developed 2 models using supervised machine learning. Model 1 (mask wearing classifier) is a text classifier for determining whether a tweet is related to mask wearing in the COVID-19 context; and Model 2 (opinion expression classifier) determines whether a tweet contained personal opinions. To train these models, we annotated a total of 1000 tweets through the following 2 steps. First, 2 authors (L.H. and C.H.) separately coded a random sample of 200 tweets to calibrate the annotation. The interrater agreement ratio was 0.93; differences were resolved in consensus development meetings. Then, the same 2 authors independently coded an additional random set of 800 tweets. To prepare for text classification, all tweets were pre-processed by, eg, lowercasing, removal of punctuations and hashtags, and stemming. Next, for each model, words were represented as 100-dimensional vectors trained by the word2vec algorithm [138]. To address the issue of imbalance between relevant and irrelevant tweets in the training data, we used the oversampling strategy as specified by Hilario et al [98]. Then, we tested several commonly used machine learning models including support vector machine, XGBoost, and long short-term memory (LSTM) network. The best-performing one, LSTM, was selected for further analyses, which achieved the highest F1-score under 10-fold cross validation. LSTM, due to its ability to account for sequential information and order dependencies, is particularly suited for handling data such as time series and natural language. Extant literature has demonstrated the predictive power of LSTM on natural language processing (NLP) tasks such as document classification and sentiment analysis [204]. We added a dropout layer (rate = 0.3) to avoid overfitting, a common technique used in the literature that randomly removes units of neutral networks [81].

To identify tweets posted by social bots (ie, programmed Twitter accounts that generate posts automatically) [74], we first applied Botometer [4, 214], a well-established social bot detection

tool that has been widely used by researchers and organizations such as the Pew Research Center [45]. Following Rauchfleisch and Kaiser's recommendation, we used a random sample of 500 distinct users from our dataset to validate the tool's performance [160]. Two authors (L.H. and C.H.) manually reviewed these users' profiles (eg, profile picture, description, account creation time, number of followers), tweeting history (eg, number of tweets, retweets, and likes), and interactions with other users (eg, commenting on others' tweets), in order to determine if a user was a social bot or not. The interrater reliability is 100% when calibrating based on 100 users' data; none of these users were determined as social bots. We then annotated the remaining 400 users; none of them were determined as social bots either. On the contrary, Botometer labeled 29 (5.8%) users as social bots. However, based on manual review, many of these users were simply hyperactive tweeters. Their online activities did exhibit the normal behavior of human users, eg, the content that they posted did not appear to be automatically authored and they participated in active interactions with other Twitter users. Removing these users could thus result in systematic biases in our analysis of the data.

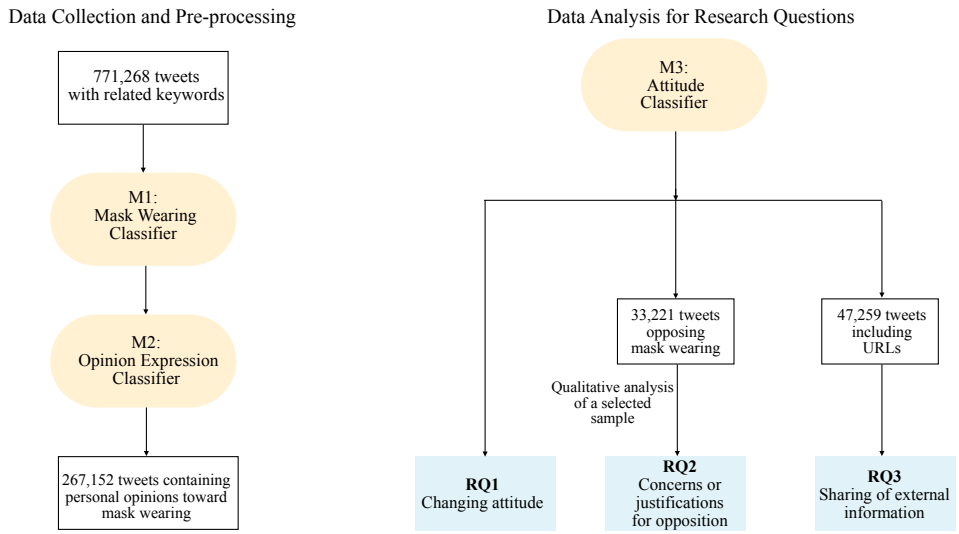

Figure 3.4: Method flowchart. RQ: research question.

Data analyses

RQ1a: Attitude

To classify public opinions toward mask wearing, we first applied the sentiment analysis approach which has been commonly used in the literature to study attitudes expressed in social media data [93]. We tested 4 off-the-shelf sentiment analysis tools that have been most commonly used: VADER (Valence Aware Dictionary for sEntiment Reasoning) [107], TextBlob [13], Stanford NLP [133], and Linguistic Inquiry and Word Count (LIWC) [188]. We manually annotated the sentiment of 500 random tweets and compared the results to the outputs of these tools. We found that none of these off-the-shelf tools was able to produce

accurate attitude classifications, at least not in our study context. The results produced by VADER, TextBlob, Stanford NLP, and LIWC achieved low F1 scores of 59%, 57.4%, 58.6%, and 51.9%, respectively. This is likely because of domain transfer- ability issues, ie, such tools are often trained with text corpora from non-healthcare domains such as movie reviews. Also, in our study, positivity of the sentiment is often not consistent with the mask wearing attitude expressed. For example, users tend to use strong negative tones such as "*Fuck. Mask On!*" to encourage others to wear masks, the sentiment of which was labeled by the off-the-shelf sentiment analysis as negative, even though the tweet was in fact in favor of mask wearing.

Thus, we developed a specialized machine learning model (model 3 [attitude classifier]) instead, using the same approach adopted in model 1 and model 2. This produced an LSTM text classifier for attitudes, trained based on 500 annotated tweets.

RQ1b: Evolution of attitude

Further, to investigate the temporal trend of public attitudes toward mask wearing, we grouped tweets by week and calculated the per- centage of tweets expressing support for, or opposition to, mask wearing on a week-to-week basis. We also analyzed word frequencies based on term frequency-inverse document frequency to assess changes in commonly discussed topics related to mask wearing over time.

RQ2: Concerns or justification for opposition

We conducted a manual qualitative content analysis on a random set of tweets posted by distinct users in order to answer RQ2 ("among those expressing an anti-mask opinion, what are their concerns or justifications"). We did not opt to use a computational approach for this RQ because the concerns expressed in tweets were heterogeneous and subtle that were challenging for the machine to classify. For example, the following tweet, "*Where I live in Los Angeles—You can't get groceries without a mask... In reality, very few are sick in CA. 250 out of 40 million have died, most with preexisting health issues*," implies that coronavirus

does not cause many casualties and masks are not necessary; hence, the attitude expressed in this tweet was opposing mask wearing. This level of natural language understanding is difficult for currently available lexicon- based tools or machine learning models to disentangle, especially on short texts such as tweets [187]. We therefore decided to qualitatively analyze such concerns and justifications for opposing mask wearing. We used grounded theory to code the data [55]. Using this method, we first randomly selected 100 tweets for open coding to generate a set of initial codes (eg, perceived physical harm and discomfort). Differences were resolved through consensus development research meetings, which produced a final set of codes for coding the rest of the data. During the open coding, saturation was achieved after coding approximately 70 tweets.

Based on the finalized codebook (provided in Appendix A), 2 authors (L.H. and C.H.) independently coded 100 randomly selected tweets to calibrate coding. The interrater reliability was 0.87. Then, each of them separately coded an additional set of 200 randomly selected tweets. Thus, in total, 500 tweets were coded and analyzed to answer this RQ.

RQ3: Sharing of external information

In this analysis, we extracted all external URLs embedded in the tweets (eg, websites, images, or videos). We then calculated and compared the proportion of pro- and anti-mask tweets that cited external sources of information. Then, we analyzed the nature of such external information, for example, whether the information originating from public health authorities such as the WHO, the CDC, and state- or county-level health departments (based on the URLs [eg, "cdc," "who," ".gov", ".int"]) was cited differently between the pro- and anti-mask groups.

We further did a drill-down analysis through manually reviewing a random set of 100 anti-mask tweets that contained external links. The objective was to specifically investigate what types of external information was used to support the anti-mask attitude. To do this, we read each of these tweets and followed the external links to review and analyze the source

information cited (eg, news articles, journal papers, and videos and images).

3.2.4 Results

A total of 771,268 tweets met our inclusion criteria. Model 1 (personal mask wearing classifier) achieved an F1 score of 84.48% under 10-fold cross validation. After applying this model, 463,369 tweets that were not relevant to mask wearing were removed. The remaining 307,899 tweets were then analyzed using model 2 (opinion expression classifier), the purpose of which was to exclude tweets that did not express a personal opinion, or the opinion was difficult to discern. Model 2 achieved an F1 score of 86.38% under 10-fold cross validation. This model further removed 40,747 irrelevant tweets, leaving a total of 267 152 tweets used in the subsequent model 3 (attitude classifier) analysis. Model 3 achieved an F1 score of 90.16% under 10-fold cross validation.

Descriptive analysis

Figure 3.5 exhibits the temporal trend of relevant tweets from January 1 to November 1, 2020. Several distinct phases can be observed. Phase I started on February 27, around the time when a statement was issued that the CDC "*does not currently recommend the use of face masks*," [41] and ended around March 30, representing a gradual increase in the volume of relevant Twitter discussions. Phase II lasted until around June 1, representing continued public interests in the subject with some fluctuations in tweet volume. A sharp increase of the number of relevant tweets followed, starting around June 2 and lasting approximately 1 month until July 6 (phase III), which may be associated with public reactions toward facial mask mandates by large states such as California. This phase was followed by another dramatic increase of relevant tweet volume (phase IV), leading toward a second peak around July 13, before returning to the phase II level, around July 20. The last phase, phase V, showed a downward trend of relevant tweet volume until the end of October, when the data

collection for this study stopped.

Table 3.6 reports the results of a word frequency analysis for examining the evolution of commonly discussed topics across these 5 distinct phases. Initially, in phase I (2/27-3/30), the discussions focused on symptoms of COVID-19 and its mechanisms of transmission. In phase II (3/31-6/1), frequently appearing topics included daily experience coping with the pandemic and the reactions to the recommendation of wearing facial masks by the CDC. Frustration and distress can be observed in phase III (6/2-7/6), expressing strong sentiments toward mask wearing. This could be associated with many circumstantial factors such as isolation, mask mandates, and a series of protests that broke out across the United States. The facial mask discussions continued in phase IV (7/7-7/20), and in phase V (7/21-10/31), moved on to focus on school reopening and the U.S. general election.

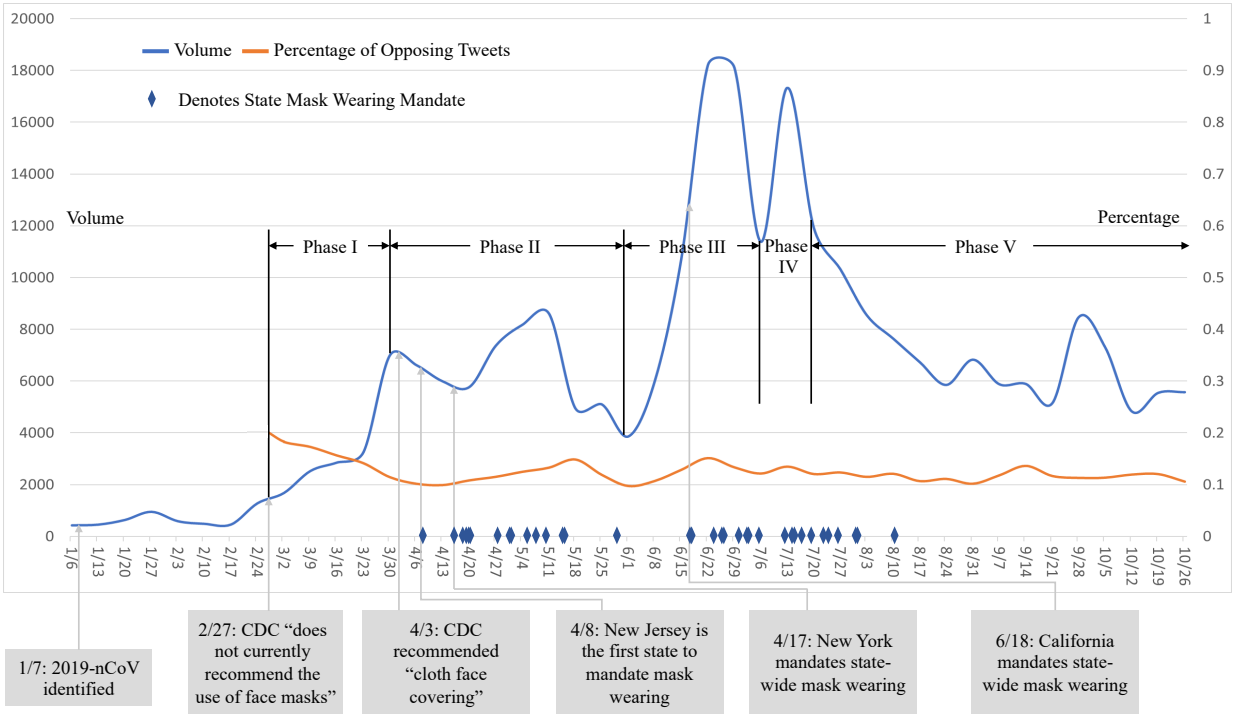

Figure 3.5: Temporal change of tweet volume and attitude. CDC: Centers for Disease Control and Prevention.

Phase	Frequently observed words	Topic
Phase I (2/27–3/30)	symptoms, coughing, sneeze	Symptom
	droplets, mouth, touch, skin, airborne, respiratory, spreading	Mechanism of transmission
	protection, sanitizer, hand, washing, soap, water, sleep, eye, face, wash, clean, mask, scarf	Best practices for personal protection
Phase II (3/31–6/1)	walking, breath, grocery, outside, seeing, wear, people, place, public, stores, shopping, car, talking, shop	Daily experience during the pandemic
	CDC, cloth, mask, bandana, mask, wearing, apart, home, covering, homemade, protect, quarantine	CDC recommendation on facial mask wearing
Phase III (6/2–7/6)	fucking, dumb, stupid, wrong, bad, fuck, hate, damn, hard, selfish, lives, risk	Distress with strong emotions
Phase IV (7/7–7/20)	people, asthma, mask, mandate, science, enough, children, folks, mandatory, life, rights, never, shut, refuse	Continued discussion on a variety of topics
Phase V (7/21–10/31)	kids, family, wear, school, safe, first, day	School reopening
	vote, trump, tested, distancing	US general election

Table 3.6: Evolution of frequently discussed topics over time

RQ1a: Attitude

Overall, during the 10-month period, 87.56% of the relevant tweets (233,931) expressed a supportive attitude toward mask wearing, whereas 12.44% (33,221) opposed it; the latter is shown as the orange line in Figure 3.5. This finding confirms our hypothesis that pro- mask tweets would outnumber anti-mask tweets; although the latter were not uncommon. Note that in Figure 3.5, we only plotted data from February 27 onward, as the number of relevant tweets at the initial stage of the pandemic was very small.

RQ1b: Evolution of attitude

As Figure 3.5 shows, initially, when the CDC recommended against the use of facial masks, the proportion of opposing tweets was around 20%. This was followed by a gradual decrease, and around the time when the CDC began to recommend mask wearing (April 3), the proportion of opposing tweets had dropped to about 10% level. Between February and October, there were a number of bursts in the volume of opposing tweets, all of which appear to be associated with state-level mask mandates. The overall opposing rate nonetheless remained around 10% to 15% throughout the study period.

RQ2: Concerns and justifications for opposition

About one-fifth of the relevant tweets (17.69%) included external information (eg, websites, images, or videos) to support the pro- or anti-mask arguments. Commonly used sources of external information included other tweets (31,619), Instagram images (10,742), YouTube videos (705), the CDC website (189), The New York Times (134), and CNN.com (134). Table 4 shows a comparison between the pro- and the anti-mask groups. Those opposed mask wearing appear to be much less likely to include external information in their tweets, compared with those supporting. This difference is statistically significant ($p < .05$). Further, the opposing group was statistically less likely to use external information from public health authorities such as the CDC, WHO, and other state- or county-level health departments ($p < .05$). These findings confirm our hypothesis that anti-mask tweets would be less likely to cite external sources of information especially from public health authorities.

Based on the qualitative content analysis of a sample of opposing tweets revealed 6 major categories of concerns or justifications for opposing facial masks: (1) physical discomfort or negative effects (30.6%), (2) lack of effectiveness (27.4%), (3) unnecessary or inap- propriate for certain people or under certain circumstances (17%), (4) political beliefs (12.2%), (5) lack of mask-wearing culture (9.6%), and (6) coronavirus not a serious threat (3.2%). Table 3.7 provides more details on each of these categories.

As shown in Table 3.7, the most common concerns expressed in anti-mask tweets are related to "*physical discomfort and negative effects*" (30.6%). These included difficulties in breathing and causing sweaty face and foggy glasses or skin ailments (eg, rash and acne). A substantial proportion of the opposing tweets also argued that facial masks were not effective in preventing the spread of coronavirus (27.4%), because the particles carrying the virus were too small and ordinary cloth coverings would not be able to stop them from penetrating through. The next notable category is beliefs that facial masks were "*unnecessary or inappropriate for certain people or in certain situations*" (17%), believing that only those infected with

the virus needed to wear masks, or that masks were only useful in settings in which social distancing was not possible. Additional reasons for the anti-mask attitude included "*political beliefs*" (12.2%), postulating that mask mandates were unconstitutional that infringed upon one's personal liberty, and were primarily politicians or the government's attempt to control the thinking and behavior of the people; "*lack of mask-wearing culture*" (9.6%), eg, wearing a mask is associated with panic and fear; and "*coronavirus not a serious threat*" (3.2%), which involved beliefs that the threat of coronavirus had been intentionally overstated.

Category	Description	Example	Proportion (N=500)
Physical discomfort or negative effects	Perception or experience of discomfort or negative effects as a result of mask wearing such as rash, acne, shortness of breath, or fainting; or beliefs that wearing a mask would cause damage to the immune system.	"It is mandatory to wear a mask at work at my very physical job will cause restrictions to airflow, making it tough to breathe. Is CDC correct or city?!!!! WTF"	30.6%
Lack of effectiveness	Beliefs that mask wearing is not effective as it claims to be, or is not always effective (e.g., if not properly worn), or there are other better alternatives.	"Non-medical face mask made of clothes are not useful for COVID-19 – NAFDAC warns URL"	27.4%
Unnecessary or inappropriate for certain people or under certain circumstances	Beliefs that healthy individuals, children, and/or those with certain health conditions should not wear masks, or that masks are not necessary outdoors or when social distancing is practiced.	"If you are not sick, you don't need to wear a mask.... people are so dumb."	17%
Political beliefs	Beliefs that mandatory mask-wearing policies infringe upon personal liberty, or that those mask mandates are politicized and are manipulation tactics by certain politicians and special interest groups	"That's because we don't want to be forced to wear a mask!!!"	12.2%
Lack of mask-wearing culture	The negative connotations associated with mask wearing such as being odd-looking, "unAmerican," criminal resembling, or reflective of panic and fear.	"I have always been made uncomfortable by someone wearing a surgical mask. Masks make me feel uneasy. This alone is enough to keep me inside. Well done, CDC."	9.6%
Coronavirus not a serious threat	Coronavirus is not a serious threat, or not as serious as what the government suggests, and thus widespread mask wearing is an overreaction.	"During flu season you run into many people who have been exposed to the flu without knowing it and in turn expose you and we are not wearing a mask all flu season! Coronavirus is not a new virus, it is just a new strand!"	3.2%

Table 3.7: Major categories of concerns or justifications for opposing mask wearing (examples paraphrased to protect confidentiality)

RQ3: Sharing of external information

About one-fifth of the relevant tweets (17.69%) included external information (eg, websites, images, or videos) to support the pro- or anti-mask arguments. Commonly used sources of external information included other tweets (31,619), Instagram images (10,742), YouTube videos (705), the CDC website (189), The New York Times (134), and CNN.com (134). Table 3.8 shows a comparison between the pro- and the anti-mask groups. Those opposed mask wearing appear to be much less likely to include external information in their tweets, compared with those supporting. This difference is statistically significant ($p < .05$). Further, the opposing group was statistically less likely to use external information from public health authorities such as the CDC, WHO, and other state- or county-level health departments ($p < .05$). These findings confirm our hypothesis that anti-mask tweets would be less likely to cite external sources of information especially from public health authorities.

Based on the qualitative content analysis that we conducted on a subset of the relevant tweets, we performed a drill-down analysis of the external supporting evidence cited in tweets that opposed mask wearing. The results show that the opposing tweets often cited user-created YouTube videos arguing against mask wearing (eg, "COVID-19 MASK REFURBISHMENT," [2]) and news articles from conservative news media (eg, Newsmax). While opposing tweets were less likely to cite information disseminated by public health authorities, a few articles published in prestigious medical journals were prominently featured, such as the perspective article "Universal Masking in Hospitals in the Covid-19 Era" published on May 21 in The New England Journal of Medicine stating that, "*We know that wearing a mask outside health care facilities offers little, if any, protection from infection.*" [15] Further, the opposing tweets commonly referenced health experts within their social circles to support the anti-mask arguments, eg, "*A family member of mine is a dr of infectious diseases. Surgical mask is 8-10 microns a n95 mask is 4 microns filter size the covid virius is .6 - 1 micron meaning the mask filter is 4-10x bigger then the virius is Covid can enter the body through your eyes and ears But we dont cover those why?*"

Source	Pro-Mask Tweets (%)	Anti-Mask Tweets (%)
Other tweets	12.96	3.88
Instagram	4.48	0.75
YouTube	0.28	0.16
CDC website	0.073	0.054
The New York Times	0.055	0.039
Websites of local public health agencies (e.g., coronavirus.ohio.gov)	0.093	0.03
Information sourced from any public health authority	0.178	0.093
Total	19.35	5.98

Table 3.8: Comparison of use of external information among pro- vs anti-mask tweets

3.2.5 Discussion

In this study, we analyzed public opinions expressed on Twitter regarding whether to or not to wear facial masks to help to prevent the spread of coronavirus. We studied a total of 771,268 tweets collected from January 1 to November 1, 2020 using an analytical strategy that combined qualitative content analysis for understanding opinions expressed in subtle human language and machine learning for scalability. The results show that while the overall volume of mask-related tweets fluctuated according to real-world events (eg, WHO/CDC recommendations and mask mandates), the proportion of anti-mask tweets stayed constant at approximately 10%. The top 3 reasons for opposing public mask wearing were physical discomfort and negative effects, lack of effectiveness, and being unnecessary or inappropriate for certain people or under certain circumstances. The results also show that anti-mask tweets were significantly less likely to use external sources of information to support the arguments, particularly information from public health authorities such as WHO and the CDC.

While there has been a large body of literature analyzing social media data to understand public opinions toward controversial health-related issues, many studies simply applied off-the-shelf sentiment analyzers by equaling the sentiment of an expression to the attitude expressed in the expression [69, 60], which may lead to incorrect interpretation of the data. Such studies may also suffer from poor domain transferability of the existing sentiment analysis tools, as most of them were developed in nonhealth domains (eg, movie reviews) [93]. In this study, instead of using the off-the-shelf tools, we developed a comprehensive computational pipeline that included multiple machine learning models trained on human-annotated data. These models helped us improve the relevance of data retrieved by keywords search by excluding tweets that were not related to the mask-wearing behavior or did not contain personal opinions. These models also helped us achieve more accurate classification of pro- or anti-mask attitude expressed.

To the best of our knowledge, no studies to date have used social media data to investigate the public's attitude toward mask wearing, except for 1 medRxiv article that used unsupervised topic modeling to look at the general discussion trends related to use of facial masks [167]. Conventional surveys available only reported on respond- ents' pro- vs anti-mask stances [114, 41, 46]. In contrast, our study was able to understand the reasoning of those who were against mask wearing. Further, our study was able to capture the evolution of attitudes over time to reveal the public's reactions to the constantly changing public health recommendations and local, regional, and national mask mandating policies. Understanding such longitudinal trends can be cost-prohibitive to achieve using the conventional survey method. Despite the advantages, use of social media data such as tweets has several limitations. First, the only way to reliably identify tweets posted by users in the United States is to use geocodes, which may introduce self-selection bias as not all Twitter users would opt to turn on geotracking. Further, the dominance of active/vocal users on Twitter, and on any other social media platforms more generally, may introduce additional self-selection biases in the data.

The results of this study provide several implications for researchers, public health practitioners, and policymakers. First, methodologically, we noticed that how to properly retrieve tweets relevant to mask wearing in the context of the pandemic requires careful consideration. Initially, we followed the common practice used in prior research (eg, Sanders et al) [167] in searching for such tweets by including COVID-19–related keywords in addition to mask-related keywords. We found that doing this would result in a loss of nearly 50% of the relevant data, owing to the fact that many tweets relevant to COVID-19 did not explicitly use any word related to COVID-19 because of its prevalence in public discourse. Further, as mentioned earlier, we found that commonly used off-the-shelf sentiment analyzers (eg, VADER, LIWC, Stanford NLP, TextBlob) failed to produce accurate sentiment classifications, or the sentiments identified were not in accordance with the attitudes expressed. Therefore, we suggest that future studies thoroughly compare existing computational tools and, if needed, train specialized, domain-specific tools to ensure the validity of study results.

Second, our data analysis revealed several distinct phases of mask-related discussions on Twitter, which closely aligned with real-world events such as shifting recommendations from public health authorities, mask mandates issued at the state level, and other contemporary events such as school reopening and the U.S. general election. This finding confirms previous studies [69, 87] and suggests that social media can be used as a reliable source of information for continuously monitoring the changing public attitude in response to major social, political, and public health events. This may provide implications into designing and implementing a social media–based real-time dashboard to monitor and track public opinions toward important health-related issues, so that health communication strategies could be more targeted and thus effective.

Third, our analysis also revealed common reasons underlying the anti-mask opinions, eg, perceived physical damage and discomfort. While there was information addressing some concerns on public health authority websites [42] and disseminated through traditional news

media [42], it appears that such information had not become highly visible through social media. Therefore, public health authorities may consider finding more creative and engaging ways to provide public education and combat with misinformation on social media platforms, eg, by creating entertaining YouTube or TikTok short videos to show how to make facial masks and how to properly wear them [101].

Further, some users opposed mask wearing because they believed that it was not an effective measure for preventing the spread of the virus and it was unnecessary for healthy individuals. These stances were indeed supported by authoritative bodies such as the CDC and the U.S. Surgeon General at the early stage of the pandemic. However, many of these tweets were posted in June and July, yet the CDC had reversed its position on public mask wearing in April. This finding indicates that certain segments of the U.S. population might not be aware of, or refuse to believe in, new recommendations from the CDC, and that earlier, contradicting recommendations might have a long-lasting impact. This means that public health experts and officials must clearly communicate about scientific uncertainty, reasons for specific recommendations, and the possibility that recommendations could change as more evidence emerges. In addition, public health experts should continue to engage the public so that new scientific findings, conclusions, and recommendations are immediately delivered to all members of the public.

Last, it is also interesting to note that while political beliefs are thought to be the key reason for anti-mask opinions [114], based on our analysis, it was not the most often cited reason in the anti-mask tweets. Instead, such tweets emphasized physical discomfort and negative effects and lack of effectiveness. However, based on our data, we are unable to determine if these reasons had ties to political beliefs that might not be made explicit in these tweets.

The findings of this study provide several insights into developing better public health communication strategies to convey the benefits of wearing facial masks, using other protective measures, as well as combating with misinformation. First, transparency is paramount in

public heath communication on sensitive and/or controversial issues [78]. This is particularly true for politically charged debates such as mask wearing. As our data show, misinformation about the lack of effectiveness of using facial masks was widespread, and many opposing opinions reflected strong political beliefs that mandatory mask wearing had been used as a tool by certain political groups to control and manipulate the public. For example, recently, the state officials of California refused to disclose the data and reasoning for the state's decision to lift the stay-at-home order, stating that "they rely on a very complex set of measurements that would confuse and potentially mislead the public if they were made public." [146] This lack of transparency, intentionally or unintentionally, will likely lead to loss of public trust and incubate conspiracy theories. Second, given the complexity, the efficacy of community mask wearing on preventing the spread of SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) is difficult to definitively prove using scientific methods. As a result, the general public tends to cherry pick the results aligned with their beliefs, or they use cautionary language commonly used in study limitation sections, eg, "the results of this study may not be generalizable," as proof of lack of evidence. As our data show, some sentences from a handful of studies published in high-impact journals (eg, The New England Journal of Medicine) were taken out of context and shared widely among those holding anti-mask beliefs. Therefore, in trying times such as this, public health officials need to put an extraordinary effort in educating the public how to properly interpret the findings reported in scientific studies [85], in addition to proactively addressing misinformation that may result from inconclusive research findings. Third, policy changes are often inevitable as new scientific evidence emerges. The rationale of such changes must be well articulated to the public. Public health officials should also not be hesitant to admit the mistakes that they might have made at the early stage of this unprecedented global pandemic due to the lack of information and uncertainties in decision making. Last, public health agencies should develop a means to constantly monitor the public's opinions, particularly those circulated through social media, in order to timely adjust their communication strategies in response

to viral spread of misinformation, misinterpretation, or misbelieves.

Future work should develop more effective machine learning classifiers to facilitate opinion mining using social media data, tweets in particular, which are often short and informal, so that automatic and continuous extraction and monitoring of public opinions are possible. In addition, future work should include more diverse social media platforms representing different types of user groups and different interaction modalitities and use qualitative approaches such as interviews and focus groups to obtain a more in-depth understanding of why certain segments of the population have a strong attitude against mask wearing, rather than relying solely on their publicly available social media posts.

3.2.6 Conclusion

Public mask wearing, while believed to be an essential personal protection measure to contain the COVID-19 pandemic, has provoked significant controversies in the United States. Through an analysis of a large Twitter dataset using a combination of qualitative content analysis and machine learning approaches, this study classified the public's attitude toward mask wearing and the evolution of this attitude over time. The results show that while most tweets were pro- mask, opposing opinions were not uncommon, and the proportion stayed rather constant throughout the pandemic to date. Common reasons for the anti-mask attitude included physical discomfort and negative effects, lack of effectiveness, and being unnecessary or inappropriate for certain people or under certain circumstances. Based on these findings, we recommend public health agencies improve their communication strategies to better convey to the public the benefits of mask wearing and combat with misinformation. Such strategies may include increased transparency in data and reasoning, being not afraid of admitting mistakes that might have been made at the early stage of the pandemic due to the lack of information, and educating the public on how to properly interpret inconclusive

or conflicting findings from scientific studies.

Chapter 4

Study 3: Developing and validating a Natural Language Processing (NLP) pipeline for clinical information extraction from notes of veterans with lymphoid malignancies

4.1 Study summary

Because electronic health records (EHR) data are often incomplete and inaccurate, they can be greatly complemented by unstructured clinical notes. Natural language processing (NLP) techniques are used to extract information from clinical notes to facilitate decision making and research. However, they are less established for rare diseases such as lymphoid malignancies due to the lack of annotated data as well as the heterogeneity and complexity

of how clinical information is documented. In this paper, we report the development and validation of an NLP pipeline that extracts clinical information such as performance status, staging, environmental exposures, and diagnosis of different types (primary, secondary, and differential) from clinical notes of veteran patients with lymphoid malignancies. We further discuss the challenges encountered in developing and deploying the NLP pipeline on Veterans Affairs (VA) data.

4.2 Introduction

Lymphoid malignancies (LM) are rare cancers; only around 145,000 patients are diagnosed in the United States per year [10, 5, 9, 7, 8]. Among veteran patients who were deployed overseas, exposures to environmental toxins and chemical agents are risk factors associated with LM. There is a need for capturing past environmental exposures of veterans to better care for them. For example, the Promise to Address Comprehensive Toxins (PACT) Act has been passed in 2022 to expand healthcare benefits and assist clinical research for veterans who have been exposed to environmental toxins [19].

While the need for identifying and caring for veterans with environmental exposures increases, such information is often not well captured in structured Electronic Health Records (EHR). In fact, most of the clinical information that is vital to clinicians' decision making is embedded in free-text, unstructured clinical notes [164]. As it is often tedious and time-consuming for clinicians to perform manual chart review to retrieve such information, clinical Natural Language Processing (cNLP) has produced numerous tools to automate the process of extracting clinical elements from unstructured notes [206]. While cNLP is an established field with mature cNLP software such as the clinical Text Analysis and Knowledge Extraction System (cTakes) [169] and the Clinical Language Annotation, Modeling, and Processing (CLAMP) [179], there is a lack of NLP resources for extracting information

from clinical notes of patients with rare cancers [200], which may be due to the difficulty of obtaining high-quality annotation data from rare disease experts. In addition, the available cNLP pipelines are based off clinical information from patients with more prevalent cancers have distinct patterns that do not apply to the rare cancer. Therefore, there is a need to develop cNLP tools that are specifically tailored to and validated on clinical notes for rare cancers so that they can capture the unique clinical information. In this paper, we developed a rule-based cNLP pipeline that was validated on clinical notes from veterans with LM within the Veterans Affairs (VA) Healthcare Systems.

4.3 Material and methods

Data collection

We identified and included veteran patients in the VA Corporate Data Warehouse (CDW) based on their histology according to the International Classifications of Disease for Oncology (ICD-O), Third Edition [6]. The Appendix A summarizes the ICD-O codes included. As a patient's record may be associated with numerous types of notes at different time points, we included all notes of the patients regardless of note types and visit time to minimize the bias. Structured data were extracted to provide patient demographic information and compare with information extracted using the NLP pipeline. To develop the pipeline, structured data and clinical notes from 961 unique patients were extracted.

Clinical entities of interest

We extracted the following clinical entities that are known to be prognostic in the care of patients with LM but are often inconsistently documented in a structured format.

Diagnosis: primary diagnosis, secondary diagnosis, and differential diagnosis

Substance use: alcohol, drug, and tobacco use.

Environmental exposure: Agent Orange, Vietnam, shipyards, Marine Corps Base Camp Lejeune

Staging: stage, Rai staging, the Multiple Myeloma International Staging System (ISS), Ann Arbor Staging, Binet staging

Performance status: Performance Status (PS), Karnofsky Performance Status (KPS), Eastern Cooperative Oncology Group (ECOG) Performance Status

Figure 4.1 provides an overview of the development and validation of the cNLP pipeline. Two independent medical experts annotated a random sample of 100 notes. Inter-rater agreement ratio was calculated, and disagreements were discussed to finalize the annotation protocol. A final set of 287 notes was annotated, which was used as the development set (dev set) for developing the cNLP pipeline.

The cNLP was iteratively developed by examining annotated texts in the clinical notes. For example, "ECOG", "KPS", and "Performance Status" were included in the dictionary of the cNLP pipeline to locate potential mentions of performance status. Similarly, "Stage", "Stg", and "Rai" were used in the dictionary to identify mentions of staging information. "Tobacco", "smoke", "alcohol", "drug", and other commonly used terms were included to extract substance use mentions. All clinical notes were preprocessed by converting words into lowercases. Punctuations are preserved. We accounted for misspellings or format issues (e.g., extra or missing spaces) by curating the regular expressions.

Several modules are based on an established cNLP software, CLAMP. We pilot tested the performance of CLAMP on the clinical entities of interest and decided to use it as the basis for extracting Diagnosis. For the rest of the clinical entities such as Performance Status, Staging, Environmental Exposures, CLAMP was not able to provide satisfying performance and often confused these entities with laboratory tests. For Diagnosis, after using CLAMP to identify a set of potential diagnosis, we curated a list of keywords that are specific to

LM to extract Primary Diagnosis, rules to categorize Diagnosis as Differential Diagnosis if the sentence includes phrases and words implying uncertainty (e.g., "differential"). Based on feedback from the preliminary error analysis, we added special modules to account for the false positives of Staging information that is associated with diseases other than LM, for example, chronic kidney disease.

The cNLP pipeline was then refined on the dev set (287 notes) . The final cNLP pipeline was then externally evaluated on a separate test set of 131 clinical notes, which were annotated by the two medical experts. The test set of 131 clinical notes was compiled in two parts to ensure the representativeness of the data as well as include enough occurrences of the clinical entities of interest. First, a random set of 100 clinical notes were selected. After being annotated by a medical expert, only 31 out of the 100 notes contain at least one clinical entity of interest, which were included in the final test set. Since solely relying on random sampling of the notes for annotation resulted in low retrieval of notes that contain clinical entities of interest, we included another set of 100 notes that had been filtered based on common keywords including "ECOG", "Performance Status", "Agent Orange". With these two sets together, a total of 131 notes with clinical entities of interest were annotated by two medical experts. The extracted tokens were evaluated based on the lenient matching criteria used in previous work [196, 97], that is, two tokens are deemed as a match if they overlap, to account for potential annotation issues and format irregularities. For each clinical entity, the precision, recall, and F1-score were calculated and presented in Table 4.1. Precision is calculated as True Positives/(True Positives + False Positives); recall is calculated as True Positives/(True Positives + False Negatives). F1-score is (2 * precision * recall)/(precision + recall).

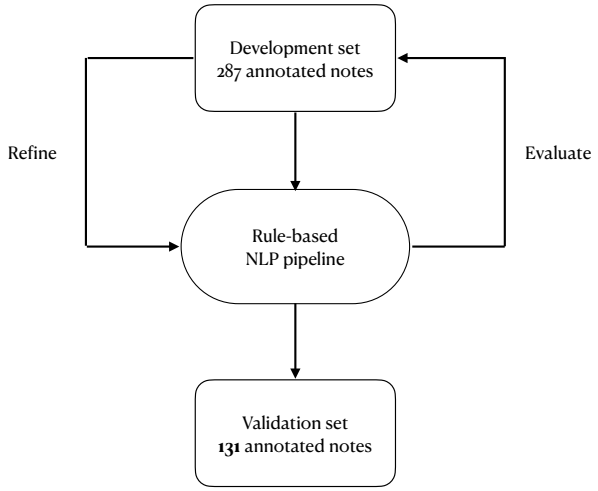

Figure 4.1: Method flowchart.

Misclassified entities in the test set were extracted and analyzed manually by the medical experts to identify potential sources of errors, in order to improve the performance of the cNLP pipeline.

This study was approved by the institutional review board (IRB) at both the VA Long Beach Healthcare System and the University of California, Irvine. Research was conducted in accordance with the Declaration of Helsinki.

Variable	Frequency in annotation	Precision	Recall	F1-score
Performance status	71	0.86	0.89	0.88
Staging	97	0.7647	0.89	0.82
Diagnosis: combined	556	0.93	0.53	0.675
Differential diagnosis	16	1	0.58	0.735
Substance use	149	0.905	1	0.95
Environmental exposures	38	0.935	0.732	0.821

Table 4.1: NLP performance on the test set with 131 notes.

4.4 Results

4.4.1 Descriptive summary

Among the 963 patients in the cohort, each patient had an average of 974 notes (standard deviation 1,071). Out of the 100 randomly selected notes, the average number of words for notes with at least one clinical entity of interest compared to notes without any clinical entities of interest was 583 and 224, respectively.

4.4.2 NLP performance

The performance of the pipeline for extracting the clinical entities on the 131 notes (test set) is reported in Table 4.1. Overall, the pipeline achieved satisfying performance, with the F1 score for all clinical entities higher than 0.6. The highest F1 score reaches 0.95 for substance use.

4.4.3 Error analysis

Lexical and linguistic variations: Some errors are attributed to lexical variations in the clinical documentation. For example, when documenting performance status, the abbreviation "PS" is commonly used in clinical notes. However, the same abbreviation is also used to indicate additional information in notes, such as postscript, and other medical diagnoses, such as pulmonary stenosis.

Lack of documentation standards: Many of the false negatives (FNs) were caused by inconsistent documentation, especially for environmental exposures. For example, while the NLP pipeline can capture common mentions such as "Agent Orange", "service in Vietnam", "chemical toxins", exposure-related information is often documented in various ways that make it difficult for the cNLP pipeline to capture comprehensively. Clinicians may use locations such as "shipyard" to imply potential exposures to toxins, while some may explicitly list the chemical toxins.

Extracted entities that are related to other diseases: Our error analysis revealed that, while some entities were identified correctly, such as staging information, they were associated with other diseases such as chronic kidney disease. Therefore, these staging-related mentions were not annotated by medical experts. These mentions comprise the majority of the false positives (FP), especially for staging information that was commonly documented for patient comorbidities. While we added an additional module to associate the disease with staging, it is not ideal to eliminate all FPs and introduce additional FNs. For example, it is common for clinicians to document the staging information for all comorbidities together (e.g., lymphoid malignancies and stage II chronic kidney disease).

Imprecise documentation: Some notes include imprecise descriptions of patients' conditions, such as "Rai staging II bordering onto III", "at least stage III". These pieces are generally not annotated by clinicians but are extracted by the cNLP pipeline as it is unable

to distinguish if the documentation is precise or vague. This type of error mainly contributes to the FPs of the pipeline.

4.5 Discussion

We developed a rule-based cNLP pipeline to identify mentions of clinical information including diagnosis, performance status, staging, substance use, and environmental exposures from clinical notes of veterans with LM. The pipeline achieved satisfying performance with F1 scores of 0.88 for Performance Status, 0.82 for Staging, 0.60 for Diagnosis, 0.735 for Differential Diagnosis, 0.95 for Substance Use, and 0.821 for Environmental Exposures. We compared the performance of the cNLP system with other systems that extracted similar clinical entities from notes of cancer patients. While we acknowledge that there are several reported cNLP systems that have higher performance (F1 of 0.955) in extracting staging and substance use information, the notes in our study sample include all types of notes for patients, while the majority of cancer-focused cNLP systems only utilize radiology and pathology notes, which are presumably more structured and exclude other clinical information unrelated to cancer [200]. The wider range of notes included in our study may bring additional challenges for the cNLP system to accurately identify clinical entities of interest due to the heterogeneity of information and documentation styles of these notes. In addition, none of the cNLP systems was designed for LM but are instead for more prevalent cancers such as breast cancer and lung cancer. Some clinical entities and their documentation are therefore inherently different from those for LM. For example, breast cancer documentation uses TNM to record staging, which is standard for most solid tumors, while LM use Ann Arbor and Rai staging [65].

Through this study, we observed several challenges that may prohibit the development of high-performing cNLP systems for LM and other rare cancers. Future efforts in developing

cNLP systems for rare cancers should consider addressing these challenges to improve the performance, alleviate medical experts' burden in annotation, and facilitate the use of cNLP systems in real-world clinical settings.

First, clinical information scatters across many different note types, which creates challenges for cNLP systems to consistently identify all of them. Previous work often focused on specific note types such as radiology reports to simplify system development, with the assumption that certain clinical information often appears in a narrow set of notes [200, 21].

Second, the characteristics of certain clinical entities make it challenging for the development of high-performing cNLP systems. For example, environmental exposure information is rarely documented in clinical notes. The sparse documentation makes it challenging to identify notes that potentially include the clinical information for medical experts to annotate. In addition, the documentation of environmental exposure information was highly inconsistent, with many mentions alluding to potential exposures through locations (e.g., shipyards, combat sites). As environmental exposure is not yet a routine component in clinical documentation, there lacks standards of consistently documenting such information.

Future directions include improving the performance of the cNLP system. To do this, we plan to leverage pre-trained large language models such as ClinicalBERT [103, 215] and CanerBERT [226] and fine-tune them on our dataset to extract the clinical entities. We will also construct a more comprehensive terminology dictionary for identifying environmental exposures for patients with LM. In addition, many clinical entities should be extracted in fine-grained forms, e.g., substance use with frequency and status, and in relation with other information, e.g., temporal information [198].

Another next step is to use the cNLP system to identify and analyze the current documentation practices and potential biases. For example, we will conduct a qualitative analysis based on the environmental exposures extracted using the cNLP system to understand the

current practices of documenting environmental exposure information for patients with LM. We will also apply the cNLP system to all veterans with LM in the VA CDW to assess the documentation patterns of the clinical information among different patient groups (by race, gender, branch, socioeconomic status), provider types, and note types to identify potential biases.

As a preliminary effort for developing a cNLP system for LM, our study has several limitations. By using a national VA HER, there is variability in data quality and completeness. As mentioned previously, elements such as environmental exposures do not have a documentation standard. The use of medical jargon and other nuances of language may also pose challenges for accurate NLP interpretation. Our development and test sets are relatively small and developed in one type of healthcare system so may not be generalizable.

4.6 Conclusion

Our cNLP study demonstrates the feasibility in leveraging advanced data analysis techniques in healthcare research, especially in extracting important clinical elements that are not complete in structured data.

Chapter 5

Discussion and future work

5.1 Discussion

5.1.1 Summary of dissertation research

In this section, I summarize both empirical and methodological contributions of my dissertation studies. I also discuss the implications from my dissertation research for future computational analysis of health text in the LLM era. Lastly, I identify several future directions that I plan to pursue as next steps.

Summary of empirical contributions

My dissertation studies provide rich empirical contributions to the health informatics community by analyzing health text data generated by patients, the general public, and clinicians. The substantive health topics that my dissertation research studied spanned across consumer health informatics, public health, and clinical informatics.

Contributions to understanding patient experiences of healthcare services

Study 2A presented in Chapter 3 provides a comprehensive patient experience taxonomy that was automatically extracted from a large-scale online physician review dataset across a ten-year period [92]. The taxonomy includes both coarse-grained topics (e.g., clinical, relationship, and management) and fine-grained aspects (e.g., treatments, communication, billing and finance). Using the taxonomy, we found that lower-rated reviews tended to mention management-related issues significantly more often than higher-rated reviews, which suggested that how healthcare providers manage their services such as processing bills and referrals and staff preparedness may have a huge impact on patients' experiences. The patient experience taxonomy can also be a valuable resource for future studies that aim to assess patient experiences with healthcare services.

Contributions to understanding public perceptions of personal protective equipment during public health crises

Study 2B described in Chapter 3 combined machine learning and qualitative analysis to analyze a large-scale twitter dataset that contained tweets related to mask wearing during the COVID-19 pandemic in the United States [91]. The study revealed the temporal trend of how the public perceived the use of masks to prevent COVID-19, which was not depicted in previous cross-sectional surveys. In addition, the study identified several public concerns toward mask wearing, including perceived lack of effectiveness and physical comfort or negative effects. These concerns, however, are not well addressed by public health authorities and even caused confusion among the public. Further, this study found that anti-mask tweets tended to share less external information compared to pro-mask tweets. Based on the study results, we provided implications for public health agencies to provide more transparent reasoning when recommending personal protective equipment such as masks during public health crises.

Contributions to understanding the limiting factors of developing NLP systems

for extracting environmental exposure information

While Study 3 reported in Chapter 4 mostly focuses on the development of an NLP system, several observations made in the study contributed to our understanding of how environmental exposure information, an underexplored social determinant of health, is documented in clinical notes of patients with lymphoid malignancies. As an initial exploration, this study found that environmental exposure information such as exposure to chemical toxins that may lead to more aggressive cancer progression is only sparsely and inconsistently documented in clinical notes. This motivates future research to more closely examine the documentation of environmental exposure information and how clinical NLP systems should be designed to more effectively extract this sparsely documented information.

Summary of methodological contributions

Best tools may still fail on your data: call for standardized pre-study evaluation of computational analysis tools on health text

The two studies presented in Chapter 2 mostly provide methodological contributions to computational analysis of health text. The most important take-away is that computational tools, even those that have been validated and achieved high performances, may still fail on health-related social media data. These tools include both rule-based software such as VADER [107] that contain carefully curated lexicons and patterns and machine learning-based software such as the Stanford NLP [133] that was trained on large-scale datasets [94, 95]. In Study 1B, I conducted a comprehensive evaluation of eleven commonly used sentiment analysis tools on five health-related social media datasets; they all performed poorly in terms of sentiment classification accuracy [94]. Further, I showed that choosing a sentiment analysis tool with poor performance can lead to biased and incorrect interpretation of the sentiment compositions of a dataset, with extreme cases where the percentage of neutral posts can be underestimated by over 200 folds when compared to ground truth labels annotated by human

coders. The tools also showed low agreement with each other when classifying sentiments in the health-related social media datasets, indicating that if researchers apply different tools on the same dataset, they may arrive at very different conclusions, which could prohibit efficient knowledge accumulation in the research community.

Study 1A revealed that the selection and use of computational sentiment analysis on health-related social media data lack consistency and rigor in the current research practices [93]. Through a systematic literature review and synthesis of the studies, I developed PATH, a protocol that encompasses comprehensive study design and reporting items for conducting computational analysis of health-related social media data and aims to improve study consistency and rigor in the health informatics community. I also provided practical recommendations for researchers such as always conducting pre-study evaluation of the tools on study data and involving multiple annotators for creating evaluation data.

Collectively, Study 1A and Study 1B warned that the health informatics community still needs caution and standardization when applying computational tools on health text and proposed PATH as a potential solution to this dire need.

No one-size-fits-all solutions: computational methods should adapt to the characteristics of health text data

Another methodological contribution of my dissertation research is a demonstration of how computational analysis can be designed and adapted to suit the characteristics of different types of health text. This is exemplified through Study 2A and 2B reported in Chapter 3. While both studies analyzed health-related social media data, the data were collected from different platforms; Study 2A collected data from Vitals.com, an online physician review website, while Study 2B retrieved Twitter data. These two datasets exhibit distinct characteristics and also unique challenges that require special computational methods for each. For example, patients and caregivers tend to mention multiple aspects of their experiences with healthcare services at the same time in a review, and these aspects are often at different

granularity. A review may only briefly compliment the clinician but criticize many parts of the clinic's management including referrals and making appointments. Therefore, the main challenge of analyzing patient-authored online physician review data is disentangling topics and aspects. To address this challenge, I designed a pipeline that first identified the coarse-grained topics through a supervised machine learning classifier and then automatically divided each topic into multiple clusters through unsupervised learning. Each cluster was reviewed by human coders to assign an aspect label such as communication, medication, and finance and billing. This pipeline efficiently used human input and produced topics and aspects at different granularity.

For Study 2B that analyzed tweets, the challenge I faced was different from Study 2A. First, I noticed that even with keyword filtering, many tweets retrieved still did not express any personal opinions toward mask wearing. Including these tweets could significantly bias our study results. Therefore, two machine learning classifiers were developed to identify tweets that did not express personal opinions (e.g., simply sharing news) and tweets that were not related to mask wearing (e.g., selling masks). Second, Twitter is known to be used by automated programs such as social bots, and particularly on heated health-related issues such as e-cigarettes, cannabis, and vaccination [22, 23]. Therefore, I also added a component in the method pipeline where Botometer, a widely used social bot detection software, was used to identify social bot accounts and eliminate their content in our study data [214].

5.1.2 Implications for computational analysis of health text in the LLM era

At the time of writing this dissertation, the advances in language technologies such as Large Language Models (LLMs) including ChatGPT brought a huge wave in both industry and academia that may reshape many fields including health informatics. Reflecting on the

research conducted before these advances, I believe many of the methods, even though deemed to be novel and promising at that time, can now be easily replaced by the emerging LLMs. This is substantiated by much recent research on the capabilities of LLMs on various tasks and datasets [228, 123, 115, 159]. Do I think the past research is meaningless? No. I believe they still hold much value now, in fact maybe even more, now that we have LLMs as powerful tools.

One of the values provided by the research in the pre-LLM era is that the methods were carefully designed and crafted based on a deep understanding of how patients, caregivers, and health providers generated different types of health texts in various scenarios. For example, the two studies presented in Chapter 4 of this dissertation employed different methods: for analyzing OPRs, the method is focused on how to effectively disentangle multiple aspects from a review; for analyzing tweets, the method devotes a large portion to pre-process the data to ensure the "purity" of the data. These study decisions were made based on manual review of sample data to understand the nature of the particular types of health text and were purposely designed to address the unique challenges (e.g., tweets often contain lots of noises and patient-generated OPRs often describe multiple aspects of their experiences at the same time).

Another key message from my research is that we should never blindly trust any computational tools, no matter how powerful they may seem. This is demonstrated through studies presented in Chapter 2. The two studies illustrated that, even computational tools specifically developed for social media data (e.g., VADER) and tools trained on large-scale datasets using advanced machine learning models (e.g., Stanford NLP) can still produce inaccurate labels on the sentiments of health-related social media data. This alerts us that any tool should be carefully validated on study data before taking the results generated by the tool as ground truths. In addition, how researchers actually use the tools is also paramount to the research community. For example, the systematic literature review reveals the highly

inconsistent practices of the use of computational sentiment analysis tools when analyzing health-related social media data. Further, researchers often don't report the details and rationals of their methods. These alerts are still applicable to the use of LLMs, as their complexities increase and how researchers use them (e.g., data pre-processing, hyperparameter selection, prompt design) can influence the validity and replicability of their research.

5.2 Future work

I see my dissertation work as a starting point rather than the end. Sitting at the intersection of many fields that are witnessing exciting advances everyday, my future research will make use of the emerging LLMs and adapt them to the healthcare context, embrace the increasingly diversified health data modality, examine potential biases in computational analysis of health data, and eventually implement data-driven systems in real-world settings to alleviate the burdens of clinicians, patients, and caregivers.

5.2.1 Integrating large language models into health text analysis

My immediate next step is to apply and evaluate various LLMs on different types of health texts including patient-generated data and clinical notes to assess their performance. For example, I plan to apply CancerBERT [226], ClinicalBERT [103], and GatorTron [215] on clinical notes from patients with lymphoid malignancies and assess whether they can perform well to extract clinical entities for rare diseases with fine-tuning. The results may provide opportunities for re-training the models and tune them to adapt to the documentation and information for rare diseases. In addition, I plan to assess the generalizability and portability of the pre-trained clinical LLMs to multiple study sites and identify factors such as how EHR system design, documentation patterns, and patient populations may affect the portability

of clinical NLP systems on downstream tasks cross different sites. The results will provide important implications for health informatics researchers and practitioners on how to adapt and customize pre-trained LLMs for their specific study sites and clinical NLP tasks.

I also plan to utilize publicly available LLMs on health-related social media data to identify patient and public attitudes and concerns toward emerging health issues. If LLMs can reliably classify patient and public attitudes with little to no annotated data (e.g., through carefully designed prompts [228] or few-shot learning), computational analysis of health-related social media data may be even more convenient in the future without requiring much annotated data from domain experts.

5.2.2 Addressing biases and disparities in computational analysis of health text

While computational analysis and LLMs afford great opportunities to unleash the potential of health text, they may also embed biases that can results in differentiated performances when applied on data from different patient populations. Extensive recent research has called our attention to the risks and consequences of biases and ethical concerns in NLP and medicine [184, 112, 106]. The biases may come from multiple sources, such as how patient information is documented in clinical notes, patient-provider interactions, etc, which are still not well understood in the field.

My immediate next step is to apply the NLP pipeline presented in Chapter 4 on all patient data and assess if the performance varies significantly across different patient groups, and if so, what factors (e.g., race, gender, provider type, etc) are associated with lower performance of the pipeline. I also plan to review practices of identifying and preventing biases introduced by computational analysis of health text and develop a checklist similar to PATH presented in Chapter 3 that can assist researchers to design and report computational analysis that

can minimize biases among different patient populations. I believe such a guideline is much needed in the health informatics community to prevent exacerbating health disparities.

Analysis of patient-generated health text may also see biases introduced by computational methods. One of my pilot studies on applying unsupervised learning on patient narratives regarding their telehealth experiences suggested that blindly applying a topic modeling algorithm may obscure patient concerns from minority groups such as Black and Latino patients. This may be due to the fact that the patient respondents were dominantly white, and topics extracted from their narratives have higher probability scores assigned by the algorithm compared to those provided by minority patient groups. This exploratory study emphasized the need to take patients' demographic information into account when conducting computational analysis of their narrative data, as well as the importance of performing subgroup analysis when the study participants are not balanced and contain minority groups whose voices may be easily obscured when merged together. I plan to continue this line of work by further exploring whether and how computational analysis may introduce biases when analyzing patient-generated narratives and identify opportunities to overcome these issues.

5.2.3 Embracing multimodal health data

While my dissertation research focuses on text data, integrating multimodal data including texts, images, videos, and audio is an emerging trend both in patient and public-generated data [216, 207] and clinical data [119]. The advances in vision-language models also enable the computational analysis of multimodal data with little to no annotated data [227, 225]. I plan to utilize multimodal health-related social media data to better understand patient and public concerns toward major health events, and how patients and the public strategically leverage multimodal data to exchange informational and emotional support. The insights from such studies may also inform public health authorities to more effectively engage the

public through interactive medias beyond text [90].

5.2.4 Implementing data-driven systems in real-world settings

Well designed systems can easily fail in real-world settings [176, 38, 26]. The failures may be attributed to policies and regulations (or the lack of),organizational issues, human factors, and technical difficulties. In the healthcare domain, such failures lead to worsened patient outcomes and even increased mortality, physician burnout, and increased healthcare expenditures. I will extend my research to incorporate implementation science and seek ways to better implement and integrate data-driven systems such as computational analysis of social media data and cNLP systems into real-world healthcare settings. This would require close collaboration with different stakeholders such as patients, caregivers, healthcare providers, and public health agencies. I have conducted a systematic review that identified the barriers and facilitators of implementing telehealth systems in low-resource settings during the COVID-19 pandemic. I will continue this line of work to explore the human, organizational, and technical factors that may affect successful implementation of data-driven systems in real-world settings.

Chapter 6

Concluding Remarks

Patients, caregivers, clinicians, and the public write and document health-related information in various forms of health text. My dissertation research, while spanned across different health-related topics, is primarily motivated by the need to develop and apply efficient computational methods to transform large-scale health text into actionable insights to improve patient and population health, assist clinicians' decision making, and facilitate clinical research. The major take-away from my dissertation research is that health text are sociotechnical products that are generated by patients, caregivers, clinicians, and the public under various restrictions (e.g., avoiding censorship of controversial health discussions by social media platforms), for different purposes (e.g., communicating personal opinions vs documenting clinical information), and shaped by technical, social, and policy factors (e.g., platform and EHR system design, culture, and healthcare policy regulations). The design and application of computational analysis methods should therefore consider and adapt to the contexts in which health text are produced.

Bibliography

[1] CAHPS Patient Experience Surveys and Guidance.

[2] COVID-19 MASK REFURBISHMENT - YouTube.

[3] General Inquirer Categories.

[4] GitHub - IUNetSci/botometer-python: A Python API for Botometer by OSoMe.

[5] Hodgkin Lymphoma Statistics | How Common Is Hodgkin Disease?

[6] ICD-O-3 | SEER Training.

[7] Key Statistics About Waldenstrom Macroglobulinemia.

[8] Key Statistics for Chronic Lymphocytic Leukemia.

[9] Key Statistics for Multiple Myeloma.

[10] Key Statistics for Non-Hodgkin Lymphoma.

[11] Lasswell Dictionary Description.

[12] SemEval.

[13] TextBlob: Simplified Text Processing — TextBlob 0.15.2 documentation.

[14] Tweet geospatial metadata.

[15] Universal Masking in Hospitals in the Covid-19 Era | NEJM.

[16] Urban Dictionary, August 24: milves.

[17] WordNet | A Lexical Database for English.

[18] N. Adams, E. E. Artigiani, and E. D. Wish. Choosing Your Platform for Social Media Drug Research and Improving Your Keyword Filter List. *Journal of Drug Issues*, 49(3):477–492, July 2019.

[19] V. Affairs. PACT Act Veterans Affairs. Accepted: 20221104.

[20] M. Afshar, S. Adelaine, F. Resnik, M. P. Mundt, J. Long, M. Leaf, T. Ampian, G. J. Wills, B. Schnapp, M. Chao, R. Brown, C. Joyce, B. Sharma, D. Dligach, E. S. Burnside, J. Mahoney, M. M. Churpek, B. W. Patterson, and F. Liao. Deployment of Real-time Natural Language Processing and Deep Learning Clinical Decision Support in the Electronic Health Record: Pipeline Implementation for an Opioid Misuse Screener in Hospitalized Adults. *JMIR Medical Informatics*, 11(1):e44977, Apr. 2023. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.

[21] P. R. Alba, A. Gao, K. M. Lee, T. Anglin-Foote, B. Robison, E. Katsoulakis, B. S. Rose, O. Efimova, J. P. Ferraro, O. V. Patterson, J. B. Shelton, S. L. Duvall, and J. A. Lynch. Ascertainment of Veterans With Metastatic Prostate Cancer in Electronic Health Records: Demonstrating the Case for Natural Language Processing. *JCO Clinical Cancer Informatics*, (5):1005–1014, Dec. 2021.

[22] J.-P. Allem, P. Escobedo, and L. Dharmapuri. Cannabis Surveillance With Twitter Data: Emerging Topics and Social Bots. *American Journal of Public Health*, 110(3):357–362, Mar. 2020.

[23] J.-P. Allem and E. Ferrara. The Importance of Debiasing Social Media Data to Better Understand E-Cigarette-Related Attitudes and Behaviors. *Journal of Medical Internet Research*, 18(8):e219, 2016. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[24] J.-P. Allem and E. Ferrara. Could Social Bots Pose a Threat to Public Health? *American Journal of Public Health*, 108(8):1005–1006, Aug. 2018.

[25] N. Alvaro, M. Conway, S. Doan, C. Lofi, J. Overington, and N. Collier. Crowdsourcing Twitter annotations to identify first-hand experiences of prescription drug use. *Journal of Biomedical Informatics*, 58:280–287, Dec. 2015.

[26] J. S. Ash, M. Berg, and E. Coiera. Some Unintended Consequences of Information Technology in Health Care: The Nature of Patient Care Information System-related Errors. *Journal of the American Medical Informatics Association : JAMIA*, 11(2):104–112, 2004.

[27] S. Baccianella, A. Esuli, and F. Sebastiani. SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. page 5.

[28] J. M. Banda, N. Adderley, W.-U.-R. Ahmed, H. AlGhoul, O. Alser, M. Alser, C. Areia, M. Cogenur, K. Fišter, S. Gombar, V. Huser, J. Jonnagaddala, L. Y. Lai, A. Leis, L. Mateu, M. A. Mayer, E. Minty, D. Morales, K. Natarajan, R. Paredes, V. S. Periyakoil, A. Prats-Uribe, E. G. Ross, G. Singh, V. Subbian, A. Vivekanantham, and D. Prieto-Alhambra. Characterization of long-term patient-reported symptoms of COVID-19: an analysis of social media data, July 2021. Pages: 2021.07.13.21260449.

[29] L. Barbera, C. Taylor, and D. Dudgeon. Why do patients with cancer visit the emergency department near the end of life? *CMAJ*, 182(6):563–568, Apr. 2010.

[30] J. Bian, Y. Zhao, R. G. Salloum, Y. Guo, M. Wang, M. Prosperi, H. Zhang, X. Du, L. J. Ramirez-Diaz, Z. He, and Y. Sun. Using Social Media Data to Understand the Impact of Promotional Information on Laypeople's Discussions: A Case Study of Lynch Syndrome. *Journal of Medical Internet Research*, 19(12), Dec. 2017.

[31] J. Blitzer, M. Dredze, and F. Pereira. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. page 8.

[32] A. Board. Why some Americans won't wear face masks, in their own words. Library Catalog: www.advisory.com.

[33] M. M. Bradley and P. J. Lang. Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. page 49.

[34] D. A. Broniatowski, A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S. C. Quinn, and M. Dredze. Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. *American Journal of Public Health*, 108(10):1378–1384, Aug. 2018. Publisher: American Public Health Association.

[35] K. Buchan, M. Filannino, and Uzuner. Automatic prediction of coronary artery disease from clinical narratives. *Journal of Biomedical Informatics*, 72:23–32, Aug. 2017.

[36] H. Bundgaard, J. S. Bundgaard, D. E. T. Raaschou-Pedersen, C. von Buchwald, T. Todsen, J. B. Norsk, M. M. Pries-Heje, C. R. Vissing, P. B. Nielsen, U. C. Winsløw, K. Fogh, R. Hasselbalch, J. H. Kristensen, A. Ringgaard, M. Porsborg Andersen, N. B. Goecke, R. Trebbien, K. Skovgaard, T. Benfield, H. Ullum, C. Torp-Pedersen, and K. Iversen. Effectiveness of Adding a Mask Recommendation to Other Public Health Measures to Prevent SARS-CoV-2 Infection in Danish Mask Wearers. *Annals of Internal Medicine*, Nov. 2020. Publisher: American College of Physicians.

[37] M. L. Cabling, J. W. Turner, A. Hurtado-de Mendoza, Y. Zhang, X. Jiang, F. Drago, and V. B. Sheppard. Sentiment Analysis of an Online Breast Cancer Support Group: Communicating about Tamoxifen. *Health Communication*, 33(9):1158–1165, 2018.

[38] E. M. Campbell, D. F. Sittig, J. S. Ash, K. P. Guappone, and R. H. Dykstra. Types of unintended consequences related to computerized provider order entry. *Journal of the American Medical Informatics Association: JAMIA*, 13(5):547–556, Oct. 2006.

[39] K. G. Card, N. Lachowsky, B. W. Hawkins, J. Jollimore, F. Baharuddin, and R. S. Hogg. Predictors of Facebook User Engagement With Health-Related Content for Gay, Bisexual, and Other Men Who Have Sex With Men: Content Analysis. *JMIR Public Health and Surveillance*, 4(2), Apr. 2018.

[40] J. Carrillo-de Albornoz, J. Rodríguez Vidal, and L. Plaza. Feature engineering for sentiment analysis in e-health forums. *PLoS ONE*, 13(11), Nov. 2018.

[41] CDC. COVID-19: Considerations for Wearing Masks | CDC.

[42] CDC. Your Guide to Masks | CDC.

[43] CDC. COVID-19 Cases, Deaths, and Trends in the US | CDC COVID Data Tracker, Mar. 2020.

[44] CDC. Transcript for CDC Telebriefing: CDC Update on Novel Coronavirus | CDC Online Newsroom | CDC, Feb. 2020.

[45] P. R. Center. How we identified bots on Twitter.

[46] P. R. Center. Most Americans say they regularly wore a mask in stores in the past month; fewer see others doing it. Library Catalog: www.pewresearch.org.

[47] P. R. Center. The Internet and Health, Feb. 2013.

[48] S. Chancellor and M. De Choudhury. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine*, 3(1):1–11, Mar. 2020. Number: 1 Publisher: Nature Publishing Group.

[49] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. page 10.

[50] Y. Chen, T. A. Lasko, Q. Mei, J. C. Denny, and H. Xu. A study of active learning methods for named entity recognition in clinical text. *Journal of Biomedical Informatics*, 58:11–18, Dec. 2015.

[51] D. Cheng, R. Kannan, S. Vempala, and G. Wang. A Divide-and-Merge Methodology for Clustering. *PODS'05*, page 26.

[52] M. Chopan, L. Sayadi, E. M. Clark, and K. Maguire. Plastic Surgery and Social Media: Examining Perceptions. *Plastic and Reconstructive Surgery*, 143(4):1259–1265, Apr. 2019.

[53] J. Chuang, J. D. Wilkerson, R. Weiss, D. Tingley, B. M. Stewart, M. E. Roberts, F. Poursabzi-Sangdeh, J. Grimmer, L. Findlater, and J. Boyd-Graber. Computer-Assisted Content Analysis: Topic Models for Exploring Multiple Subjective Interpretations. page 9.

[54] E. M. Clark, C. A. Jones, J. R. Williams, A. N. Kurti, M. C. Norotsky, C. M. Danforth, and P. S. Dodds. Vaporous Marketing: Uncovering Pervasive Electronic Cigarette Advertisements on Twitter. *PLOS ONE*, 11(7):e0157304, July 2016. Publisher: Public Library of Science.

[55] J. Corbin and A. Strauss. *Basics of qualitative research: Techniques and procedures for developing grounded theory, 3rd ed.* Basics of qualitative research: Techniques and procedures for developing grounded theory, 3rd ed. Sage Publications, Inc, Thousand Oaks, CA, US, 2008.

[56] J. Cowie and W. Lehnert. Information extraction. *Communications of the ACM*, 39(1):80–91, Jan. 1996.

[57] J. A. Cummins. Getting a Vaccine, Jab, or Vax Is More Than a Regular Expression. Comment on "COVID-19 Vaccine-Related Discussion on Twitter: Topic Modeling and Sentiment Analysis". *Journal of Medical Internet Research*, 24(2):e31978, Feb. 2022. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[58] R. Daniulaityte, L. Chen, F. R. Lamy, R. G. Carlson, K. Thirunarayan, and A. Sheth. "When 'Bad' is 'Good'": Identifying Personal Communication and Sentiment in Drug-Related Tweets. *JMIR Public Health and Surveillance*, 2(2), Oct. 2016.

[59] T. J. Daskivich, J. Houman, G. Fuller, J. T. Black, H. L. Kim, and B. Spiegel. Online physician ratings fail to predict actual performance on measures of quality, value, and peer review. *Journal of the American Medical Informatics Association*, 25(4):401–407, Apr. 2018.

[60] M. A. Davis, K. Zheng, Y. Liu, and H. Levy. Public Response to Obamacare on Twitter. *Journal of Medical Internet Research*, 19(5):e167, May 2017.

[61] A. P. Dawid. Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31, 1979.

[62] P. Delir Haghighi, Y.-B. Kang, R. Buchbinder, F. Burstein, and S. Whittle. Investigating Subjective Experience and the Influence of Weather Among Individuals With Fibromyalgia: A Content Analysis of Twitter. *JMIR Public Health and Surveillance*, 3(1), Jan. 2017.

[63] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772, Oct. 2009.

[64] K. Denecke and Y. Deng. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial Intelligence in Medicine*, 64(1):17–27, May 2015.

[65] P. R. Deshmukh and R. Phalnikar. Information extraction for prognostic stage prediction from breast cancer medical records using NLP and ML. *Medical & Biological Engineering & Computing*, 59(9):1751–1772, Sept. 2021.

[66] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. arXiv:1810.04805 [cs].

[67] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth. Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PLOS ONE*, 6(12):e26752, Dec. 2011. Publisher: Public Library of Science.

[68] J. Du, J. Xu, H. Song, X. Liu, and C. Tao. Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets. *Journal of Biomedical Semantics*, 8, Mar. 2017.

[69] J. Du, J. Xu, H.-Y. Song, and C. Tao. Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data. *BMC Medical Informatics and Decision Making*, 17(Suppl 2), July 2017.

[70] J. Dye, I. Schatz, B. Rosenberg, and S. Coleman. Constant Comparison Method: A Kaleidoscope of Data. *The Qualitative Report*, 4(1):1–10, Jan. 2000.

[71] S. E. Eikenberry, M. Mancuso, E. Iboi, T. Phan, K. Eikenberry, Y. Kuang, E. Kostelich, and A. B. Gumel. To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. *Infectious Disease Modelling*, 5:293–308, Jan. 2020.

[72] A. Esuli and F. Sebastiani. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA).

[73] G. Fernandez, C. Maione, K. Zaballa, N. Bonnici, B. H. Spitzberg, J. Carter, H. Yang, J. McKew, F. Bonora, S. S. Ghodke, C. Jin, R. De Ocampo, W. Kepner, and M.-H. Tsou. Sentiment Analysis of Social Media Response and Spatial Distribution Patterns on the COVID-19 Outbreak: The Case Study of Italy. In A. Nara and M.-H. Tsou, editors, *Empowering Human Dynamics Research with Social Media and Geospatial Data Analytics*, Human Dynamics in Smart Cities, pages 167–184. Springer International Publishing, Cham, 2021.

[74] E. Ferrara. What types of COVID-19 conspiracies are populated by Twitter bots? *First Monday*, May 2020.

[75] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, June 2016.

[76] J. Fleiss. Measuring nominal scale agreement among many raters. - PsycNET. *Psychological Bulletin*, 76(5):378–382, 1971.

[77] R. C. Ford, S. A. Bach, and M. D. Fottler. Methods of Measuring Patient Satisfaction in Health Care Organizations. *Health Care Management Review*, 22(2):74–89, Apr. 1997.

[78] P. E. French. Enhancing the Legitimacy of Local Government Pandemic Influenza Planning through Transparency and Public Engagement. *Public Administration Review*, 71(2):253–264, 2011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6210.2011.02336.x.

[79] E. Gabarron, A. Dechsling, I. Skafle, and A. Nordahl-Hansen. Discussions of Asperger Syndrome on Social Media: Content and Sentiment Analysis on Twitter. *JMIR Formative Research*, 6(3):e32752, Mar. 2022. Company: JMIR Formative Research Distributor: JMIR Formative Research Institution: JMIR Formative Research Label: JMIR Formative Research Publisher: JMIR Publications Inc., Toronto, Canada.

[80] E. Gabarron, E. Dorronzoro, O. Rivera-Romero, and R. Wynn. Diabetes on Twitter: A Sentiment Analysis. *Journal of Diabetes Science and Technology*, 13(3):439–444, May 2019.

[81] Y. Gal and Z. Ghahramani. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. *Advances in Neural Information Processing Systems*, 29:1019–1027, 2016.

[82] G. G. Gao, J. S. McCullough, R. Agarwal, and A. K. Jha. A Changing Landscape of Physician Quality Reporting: Analysis of Patients' Online Ratings of Their Physicians Over a 5-Year Period. *Journal of Medical Internet Research*, 14(1):e38, Feb. 2012.

[83] A. Goel and L. Gupta. Social Media in the Times of COVID-19. *Journal of Clinical Rheumatology*, page 10.1097/RHU.0000000000001508, June 2020.

[84] S. Gohil, S. Vuik, and A. Darzi. Sentiment Analysis of Health Care Tweets: Review of the Methods Used. *JMIR Public Health and Surveillance*, 4(2):e43, 2018.

[85] S. Gruber-Miller. Fact check: New England journal article taken out of context, didn't bash face masks.

[86] O. Gruebner, S. R. Lowe, M. Sykora, K. Shankardass, S. V. Subramanian, and S. Galea. A novel surveillance approach for disaster mental health. *PLoS ONE*, 12(7), July 2017.

[87] X. Gui, Y. Wang, Y. Kou, T. L. Reynolds, Y. Chen, Q. Mei, and K. Zheng. Understanding the Patterns of Health Information Dissemination on Social Media during the Zika Outbreak. *AMIA Annual Symposium Proceedings*, 2017:820, 2017. Publisher: American Medical Informatics Association.

[88] O. L. Haimson. Mapping gender transition sentiment patterns via social media data: toward decreasing transgender mental health disparities. *Journal of the American Medical Informatics Association*, 26(8-9):749–758, Aug. 2019.

[89] K. M. Harris. How Do Patients Choose Physicians? Evidence from a National Survey of Enrollees in Employment-Related Health Plans. *Health Services Research*, 38(2):711–732, 2003.

[90] C. He, H. Liu, L. He, T. Lu, and B. Li. More collaboration, less seriousness: Investigating new strategies for promoting youth engagement in government-generated videos during the COVID-19 pandemic in China. *Computers in Human Behavior*, 126:107019, Jan. 2022.

[91] L. He, C. He, T. L. Reynolds, Q. Bai, Y. Huang, C. Li, K. Zheng, and Y. Chen. Why do people oppose mask wearing? A comprehensive analysis of U.S. tweets during the COVID-19 pandemic. *Journal of the American Medical Informatics Association*, 28(7):1564–1573, July 2021.

[92] L. He, C. He, Y. Wang, Z. Hu, K. Zheng, and Y. Chen. What Do Patients Care About? Mining Fine-grained Patient Concerns from Online Physician Reviews Through Computer-Assisted Multi-level Qualitative Analysis. *AMIA Annual Symposium Proceedings*, 2020:544–553, Jan. 2021.

[93] L. He, T. Yin, Z. Hu, Y. Chen, D. A. Hanauer, and K. Zheng. Developing a standardized protocol for computational sentiment analysis research using health-related social media data. *Journal of the American Medical Informatics Association*, (ocaa298), Dec. 2020.

[94] L. He, T. Yin, and K. Zheng. They May Not Work! An evaluation of eleven sentiment analysis tools on seven social media datasets. *Journal of Biomedical Informatics*, 132:104142, Aug. 2022.

[95] L. He and K. Zheng. How Do General-Purpose Sentiment Analyzers Perform when Applied to Health-Related Online Social Media Data? *Studies in Health Technology and Informatics*, 264:1208–1212, Aug. 2019.

[96] health.com. Why Do Some People Refuse to Wear a Face Mask in Public? | Health.com.

[97] S. Henry, K. Buchan, M. Filannino, A. Stubbs, and O. Uzuner. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12, Jan. 2020.

[98] A. F. Hilario, S. G. López, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera. *Learning from Imbalanced Data Sets*. Springer International Publishing, 2018.

[99] T. J. Hoerger and L. Z. Howard. Search Behavior and Choice of Physician in the Market for Prenatal Care. *Medical Care*, 33(4):332–349, Apr. 1995.

[100] A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong, and U. Kaymak. Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 703–710, Coimbra, Portugal, Mar. 2013. Association for Computing Machinery.

[101] R. Hosie. Watch: Dentist on TikTok shows how to make face mask fit better.

[102] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, New York, NY, USA, Aug. 2004. Association for Computing Machinery.

[103] K. Huang, J. Altosaar, and R. Ranganath. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv:1904.05342 [cs]*, Nov. 2020. arXiv: 1904.05342.

[104] J. Huh, M. Yetisgen-Yildiz, and W. Pratt. Text Classification for Assisting Moderators in Online Health Communities. *Journal of biomedical informatics*, 46(6), Dec. 2013.

[105] J. Huppertz and P. Otto. Predicting HCAHPS scores from hospitals' social media pages: A sentiment analysis. *Health Care Management Review*, 43(4):359–367, Dec. 2018.

[106] B. Hutchinson, V. Prabhakaran, E. Denton, K. Webster, Y. Zhong, and S. Denuyl. Social Biases in NLP Models as Barriers for Persons with Disabilities, May 2020. arXiv:2005.00813 [cs].

[107] C. J. Hutto and E. Gilbert. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. page 10.

[108] T. U. o. D. Informatics and Mathematical Modelling. AFINN.

[109] J. Jagosh, J. Donald Boudreau, Y. Steinert, M. E. MacDonald, and L. Ingram. The importance of physician listening from the patients' perspective: Enhancing diagnosis, healing, and the doctor–patient relationship. *Patient Education and Counseling*, 85(3):369–374, Dec. 2011.

[110] J. Jarry. Why Some People Choose Not to Wear a Mask.

[111] K. Joseph, S. Shugars, R. Gallagher, J. Green, A. Q. Mathé, Z. An, and D. Lazer. (Mis)alignment Between Stance Expressed in Social Media Data and Public Opinion Surveys, Sept. 2021. arXiv:2109.01762 [cs].

[112] Y. J. Juhn, E. Ryu, C.-I. Wi, K. S. King, M. Malik, S. Romero-Brufau, C. Weng, S. Sohn, R. R. Sharp, and J. D. Halamka. Assessing socioeconomic bias in machine learning algorithms in health care: a case study of the HOUSES index. *Journal of the American Medical Informatics Association*, 29(7):1142–1151, July 2022.

[113] B. Kadry, L. F. Chu, B. Kadry, D. Gammas, and A. Macario. Analysis of 4999 Online Physician Ratings Indicates That Most Patients Give Physicians a Favorable Rating. *Journal of Medical Internet Research*, 13(4):e95, Nov. 2011.

[114] J. Katz, M. Sanger-Katz, and K. Quealy. A Detailed Map of Who Is Wearing Masks in the U.S. *The New York Times*, July 2020.

[115] R. A. Khan, M. Jawaid, A. R. Khan, and M. Sajjad. ChatGPT - Reshaping medical education and clinical management. *Pakistan Journal of Medical Sciences*, 39(2):605–607, 2023.

[116] A. S. Kilaru, Z. F. Meisel, B. Paciotti, Y. P. Ha, R. J. Smith, B. L. Ranard, and R. M. Merchant. What do patients say about emergency departments in online reviews? A qualitative study. *BMJ Quality & Safety*, 25(1):14–24, Jan. 2016.

[117] Y. Kim, J. Huang, and S. Emery. Garbage in, Garbage Out: Data Collection, Quality Assessment and Reporting Standards for Social Media Data Use in Health Research, Infodemiology and Digital Disease Detection. *Journal of Medical Internet Research*, 18(2):e41, 2016.

[118] K. Klimiuk, A. Czoska, K. Biernacka, and Balwicki. Vaccine misinformation on social media – topic-based content and sentiment analysis of Polish vaccine-deniers' comments on Facebook. *Human Vaccines & Immunotherapeutics*, 17(7):2026–2035, July 2021. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/21645515.2020.1850072.

[119] A. Kline, H. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, F. Wang, F. Cheng, and Y. Luo. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):1–14, Nov. 2022. Number: 1 Publisher: Nature Publishing Group.

[120] A. Lederman, R. Lederman, and K. Verspoor. Tasks as needs: reframing the paradigm of clinical natural language processing research for real-world decision support. *Journal of the American Medical Informatics Association*, 29(10):1810–1817, Oct. 2022.

[121] Q. Li, C. Wang, R. Liu, L. Wang, D. D. Zeng, and S. J. Leischow. Understanding Users' Vaping Experiences from Social Media: Initial Study Using Sentiment Opinion Summarization Techniques. *Journal of Medical Internet Research*, 20(8), Aug. 2018.

[122] B. Liu. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012.

[123] S. Liu, A. P. Wright, B. L. Patterson, J. P. Wanderer, R. W. Turer, S. D. Nelson, A. B. McCoy, D. F. Sittig, and A. Wright. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *Journal of the American Medical Informatics Association*, page ocad072, Apr. 2023.

[124] Y. Lu, X. Hu, F. Wang, S. Kumar, H. Liu, and R. Maciejewski. Visualizing Social Media Sentiment in Disaster Scenarios. In *Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion*, pages 1211–1215, Florence, Italy, 2015. ACM Press.

[125] Y. Luo, Q. Li, S. Sun, D. Zeng, and S. J. Leischow. Understanding of Users' Response to the Intervention of FDA's New Deeming Rules in Twitter. In *Proceedings of the 2nd International Conference on Medical and Health Informatics - ICMHI '18*, pages 1–8, Tsukuba, Japan, 2018. ACM Press.

[126] K. Lybarger. *Extracting Information from Clinical Text with Limited Annotated Data*. Ph.D., University of Washington, United States – Washington. ISBN: 9798684669682.

[127] K. Lybarger, M. Ostendorf, and M. Yetisgen. Annotating Social Determinants of Health Using Active Learning, and Characterizing Determinants Using Neural Event Extraction. *Journal of biomedical informatics*, 113:103631, Jan. 2021.

[128] J. C. Lyu, E. L. Han, and G. K. Luli. COVID-19 Vaccine–Related Discussion on Twitter: Topic Modeling and Sentiment Analysis. *Journal of Medical Internet Research*, 23(6):e24435, June 2021. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[129] A. López, A. Detz, N. Ratanawongsa, and U. Sarkar. What Patients Say About Their Doctors Online: A Qualitative Content Analysis. *Journal of General Internal Medicine*, 27(6):685–692, June 2012.

[130] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning Word Vectors for Sentiment Analysis. page 9.

[131] R. Mamidi, M. Miller, T. Banerjee, W. Romine, and A. Sheth. Identifying Key Topics Bearing Negative Sentiment on Twitter: Insights Concerning the 2015-2016 Zika Epidemic. *JMIR public health and surveillance*, 5(2):e11036, June 2019.

[132] L. Mamykina, D. Nakikj, and N. Elhadad. Collective Sensemaking in Online Health Forums. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3217–3226, New York, NY, USA, Apr. 2015. Association for Computing Machinery.

[133] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, 2014. Association for Computational Linguistics.

[134] L. S. Martinez, S. Hughes, E. R. Walsh-Buhi, and M.-H. Tsou. "Okay, We Get It. You Vape": An Analysis of Geocoded Content, Context, and Sentiment regarding E-Cigarettes on Twitter. *Journal of Health Communication*, 23(6):550–562, June 2018.

[135] Medhelp. MedHelp International - Products, Competitors, Financials, Employees, Headquarters Locations.

[136] S. Mehrabi, A. Krishnan, A. M. Roch, H. Schmidt, D. Li, J. Kesterson, C. Beesley, P. Dexter, M. Schmidt, M. Palakal, and H. Liu. Identification of Patients with Family History of Pancreatic Cancer - Investigation of an NLP System Portability. *Studies in health technology and informatics*, 216:604–608, 2015.

[137] C. A. Melton, O. A. Olusanya, N. Ammar, and A. Shaban-Nejad. Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence. *Journal of Infection and Public Health*, 14(10):1505–1512, Oct. 2021.

[138] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in*

Neural Information Processing Systems 26, pages 3111–3119. Curran Associates, Inc., 2013.

[139] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Annals of Internal Medicine*, 151(4):264–269, Aug. 2009. Publisher: American College of Physicians.

[140] S. A. Moorhead, D. E. Hazlett, L. Harrison, J. K. Carroll, A. Irwin, and C. Hoving. A New Dimension of Health Care: Systematic Review of the Uses, Benefits, and Limitations of Social Media for Health Communication. *Journal of Medical Internet Research*, 15(4):e85, 2013.

[141] A. Mulyar, O. Uzuner, and B. McInnes. MT-clinical BERT: scaling clinical information extraction with multitask learning. *Journal of the American Medical Informatics Association*, 28(10):2108–2115, Oct. 2021.

[142] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, 30(1):3–26, Jan. 2007. Publisher: John Benjamins.

[143] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov. SemEval-2016 Task 4: Sentiment Analysis in Twitter. *arXiv:1912.01973 [cs]*, Dec. 2019. arXiv: 1912.01973.

[144] Z. Nasar, S. W. Jaffry, and M. K. Malik. Named Entity Recognition and Relation Extraction: State-of-the-Art. *ACM Computing Surveys*, 54(1):20:1–20:39, Feb. 2021.

[145] E. National Academies of Sciences. *Rapid Expert Consultation on the Effectiveness of Fabric Masks for the COVID-19 Pandemic (April 8, 2020)*. Apr. 2020.

[146] A. B. C. News. It's a secret: California keeps key virus data from public.

[147] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, and M. Berk. Affective and Content Analysis of Online Depression Communities. *IEEE Transactions on Affective Computing*, 5(3):217–226, July 2014. Conference Name: IEEE Transactions on Affective Computing.

[148] J. S. Oh, D. He, W. Jeng, E. Mattern, and L. Bowler. Linguistic characteristics of eating disorder questions on Yahoo! Answers – content, style, and emotion. *Proceedings of the American Society for Information Science and Technology*, 50(1):1–10, 2013. _eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/meet.14505001068.

[149] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, 2:13, July 2019.

[150] N. Oscar, P. A. Fox, R. Croucher, R. Wernick, J. Keune, and K. Hooker. Machine Learning, Sentiment Analysis, and Tweets: An Examination of Alzheimer's Disease Stigma on Twitter. *The Journals of Gerontology: Series B*, 72(5):742–751, Sept. 2017.

[151] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135, 2008.

[152] P. C.-I. Pang and L. Liu. Why Do Consumers Review Doctors Online? Topic Modeling Analysis of Positive and Negative Reviews on an Online Health Community in China. page 10.

[153] N. Pawsey, T. Nayeem, and X. Huang. Use of facebook to engage water customers: A comprehensive study of current U.K. and Australian practices and trends. *Journal of Environmental Management*, 228:517–528, Dec. 2018.

[154] I. . T. Pew Research Center. Demographics of Social Media Users and Adoption in the United States | Pew Research Center, June 2019.

[155] Physionet. Responsible use of MIMIC data with online services like GPT.

[156] Y. Pruksachatkun, S. R. Pendse, and A. Sharma. Moments of Change: Analyzing Peer-Based Cognitive Support in Online Mental Health Forums. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 64:1–64:13, New York, NY, USA, 2019. ACM. event-place: Glasgow, Scotland Uk.

[157] M. Pérez-Pérez, G. Pérez-Rodríguez, F. Fdez-Riverola, and A. Lourenço. Using Twitter to Understand the Human Bowel Disease Community: Exploratory Analysis of Key Topics. *Journal of Medical Internet Research*, 21(8):e12610, Aug. 2019.

[158] M. Rastegar-Mojarad, Z. Ye, D. Wall, N. Murali, and S. Lin. Collecting and Analyzing Patient Experiences of Health Care From Social Media. *JMIR Research Protocols*, 4(3), July 2015.

[159] S. Rathje, D.-M. Mirea, I. Sucholutsky, R. Marjieh, C. Robertson, and J. J. V. Bavel. GPT is an effective tool for multilingual psychological text analysis, May 2023.

[160] A. Rauchfleisch and J. Kaiser. The False positive problem of automatic bot detection in social science research. *PLOS ONE*, 15(10):e0241045, Oct. 2020. Publisher: Public Library of Science.

[161] T. W. Reader, A. Gillespie, and J. Roberts. Patient complaints in healthcare systems: a systematic review and coding taxonomy. *BMJ Quality & Safety*, 23(8):678–689, Aug. 2014.

[162] T. L. Reynolds, N. Ali, E. McGregor, T. O'Brien, C. Longhurst, A. L. Rosenberg, S. E. Rudkin, and K. Zheng. Understanding Patient Questions about their Medical Records in an Online Health Forum: Opportunity for Patient Portal Design. *AMIA Annual Symposium Proceedings*, 2017:1468–1477, Apr. 2018.

[163] B. J. Ricard, L. A. Marsch, B. Crosier, and S. Hassanpour. Exploring the Utility of Community-Generated Social Media Content for Detecting Depression: An Analytical Study on Instagram. *Journal of Medical Internet Research*, 20(12), Dec. 2018.

[164] A. Roberts. Language, Structure, and Reuse in the Electronic Health Record. *AMA Journal of Ethics*, 19(3):281–288, Mar. 2017. Publisher: American Medical Association.

[165] M. Roccetti, A. Casari, and G. Marfia. Inside Chronic Autoimmune Disease Communities: A Social Networks Perspective to Crohn's Patient Behavior and Medical Information. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, pages 1089–1096, New York, NY, USA, 2015. ACM. event-place: Paris, France.

[166] S. T. Rosenbloom, J. C. Denny, H. Xu, N. Lorenzi, W. W. Stead, and K. B. Johnson. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association*, 18(2):181–186, Mar. 2011.

[167] A. Sanders, R. White, L. Severson, R. Ma, R. McQueen, H. C. A. Paulo, Y. Zhang, J. S. Erickson, and K. P. Bennett. Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse. *medRxiv*, page 2020.08.28.20183863, Sept. 2020. Publisher: Cold Spring Harbor Laboratory Press.

[168] A. Sarker and Y. Ge. Mining long-COVID symptoms from Reddit: characterizing post-COVID syndrome from patient reports. *JAMIA Open*, 4(3):ooab075, July 2021.

[169] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA*, 17(5):507–513, 2010.

[170] S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, and V. Osmani. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Medical Informatics*, 7(2):e12239, May 2019. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.

[171] Y. Shen, P. Clarke, I. N. Gomez-Lopez, A. B. Hill, D. M. Romero, R. Goodspeed, V. J. Berrocal, V. V. Vydiswaran, and T. C. Veinot. Using social media to assess the consumer nutrition environment: comparing Yelp reviews with a direct observation audit instrument for grocery stores. *Public Health Nutrition*, 22(2):257–264, Feb. 2019.

[172] M. Shepherd. Does The Public Response To The Latest Viral Covid-19 Cure Video Mark The Death Of Critical Thinking? Library Catalog: www.forbes.com Section: Innovation.

[173] C. Shivade, P. Raghavan, E. Fosler-Lussier, P. J. Embi, N. Elhadad, S. B. Johnson, and A. M. Lai. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2):221–230, Mar. 2014.

[174] Y. Si, J. Wang, H. Xu, and K. Roberts. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304, Nov. 2019.

[175] C. Siemaszko. Here's why some people are not wearing masks during the coronavirus crisis.

[176] D. F. Sittig, A. Wright, J. Ash, and H. Singh. New Unintended Adverse Consequences of Electronic Health Records. *Yearbook of Medical Informatics*, 25(01):7–12, Aug. 2016.

[177] S. Sohn, Y. Wang, C.-I. Wi, E. A. Krusemark, E. Ryu, M. H. Ali, Y. J. Juhn, and H. Liu. Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions. *Journal of the American Medical Informatics Association*, 25(3):353–359, Mar. 2018.

[178] S. Sohn, C.-I. Wi, Y. J. Juhn, and H. Liu. Analysis of Clinical Variations in Asthma Care Documented in Electronic Health Records Between Staff and Resident Physicians. *Studies in health technology and informatics*, 245:1170–1174, 2017.

[179] E. Soysal, J. Wang, M. Jiang, Y. Wu, S. Pakhomov, H. Liu, and H. Xu. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association: JAMIA*, Nov. 2017.

[180] E. Soysal, J. Wang, M. Jiang, Y. Wu, S. Pakhomov, H. Liu, and H. Xu. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association: JAMIA*, 25(3):331–336, Mar. 2018.

[181] M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*, EMNLP '11, pages 53–63, USA, July 2011. Association for Computational Linguistics.

[182] L. Sterckx, G. Vandewiele, I. Dehaene, O. Janssens, F. Ongenae, F. De Backere, F. De Turck, K. Roelens, J. Decruyenaere, S. Van Hoecke, and T. Demeester. Clinical information extraction for preterm birth risk prediction. *Journal of Biomedical Informatics*, 110:103544, Oct. 2020.

[183] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger. Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, 39:156–168, Apr. 2018.

[184] I. Straw and C. Callison-Burch. Artificial Intelligence in mental health and the biases of language based models. *PLOS ONE*, 15(12):e0240376, Dec. 2020. Publisher: Public Library of Science.

[185] A. P. Sunjaya and C. Jenkins. Rationale for universal face masks in public against COVID-19. *Respirology (Carlton, Vic.)*, Apr. 2020.

[186] H. Talpada, M. N. Halgamuge, and N. Tran Quoc Vinh. An Analysis on Use of Deep Learning and Lexical-Semantic Based Sentiment Analysis Method on Twitter Data to Understand the Demographic Trend of Telemedicine. In *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–9, Oct. 2019. ISSN: 2164-2508.

[187] J. Tang, Z. Meng, X. Nguyen, Q. Mei, and M. Zhang. Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis. page 9.

[188] Y. R. Tausczik and J. W. Pennebaker. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54, Mar. 2010.

[189] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[190] M. Thelwall and S. Thelwall. Retweeting for COVID-19: Consensus building, information sharing, dissent, and lockdown life. *arXiv:2004.02793 [cs]*, May 2020. arXiv: 2004.02793.

[191] P. J. Tighe, R. C. Goldsmith, M. Gravenstein, H. R. Bernard, and R. B. Fillingim. The Painful Tweet: Text, Sentiment, and Community Structure Analyses of Tweets Pertaining to Pain. *Journal of Medical Internet Research*, 17(4):e84, Apr. 2015.

[192] C. Tong, D. Margolin, R. Chunara, J. Niederdeppe, T. Taylor, N. Dunbar, and A. J. King. Search Term Identification Methods for Computational Health Communication: Word Embedding and Network Approach for Health Content on YouTube. *JMIR Medical Informatics*, 10(8):e37862, Aug. 2022. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.

[193] S.-F. Tsao, H. Chen, T. Tisseverasinghe, Y. Yang, L. Li, and Z. A. Butt. What social media told us in the time of COVID-19: a scoping review. *The Lancet Digital Health*, 3(3):e175–e194, Mar. 2021. Publisher: Elsevier.

[194] Z. Tufekci. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):505–514, May 2014. Number: 1.

[195] Twitter. Twitter API Standard v1.1.

[196] J. Vasilakes, A. Bompelli, J. R. Bishop, T. J. Adam, O. Bodenreider, and R. Zhang. Assessing the enrichment of dietary supplement coverage in the Unified Medical Language System. *Journal of the American Medical Informatics Association*, 27(10):1547–1555, Oct. 2020.

[197] S. Velupillai, H. Suominen, M. Liakata, A. Roberts, A. D. Shah, K. Morley, D. Osborn, J. Hayes, R. Stewart, J. Downs, W. Chapman, and R. Dutta. Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances. *Journal of Biomedical Informatics*, 88:11–19, Dec. 2018.

[198] N. Viani, R. Botelle, J. Kerwin, L. Yin, R. Patel, R. Stewart, and S. Velupillai. A natural language processing approach for identifying temporal disease onset information from mental healthcare text. *Scientific Reports*, 11(1):757, Jan. 2021. Number: 1 Publisher: Nature Publishing Group.

[199] B. C. Wallace, M. J. Paul, U. Sarkar, T. A. Trikalinos, and M. Dredze. A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *Journal of the American Medical Informatics Association : JAMIA*, 21(6):1098–1103, Nov. 2014.

[200] L. Wang, S. Fu, A. Wen, X. Ruan, H. He, S. Liu, S. Moon, M. Mai, I. B. Riaz, N. Wang, P. Yang, H. Xu, J. L. Warner, and H. Liu. Assessment of Electronic Health Record for Cancer Research and Patient Care Through a Scoping Review of Cancer Natural Language Processing. *JCO Clinical Cancer Informatics*, 6:e2200006, Aug. 2022.

[201] T. Wang, M. Brede, A. Ianni, and E. Mentzakis. Social interactions in online eating disorder communities: A network perspective. *PLOS ONE*, 13(7):e0200800, July 2018.

[202] X. Wang, A. High, X. Wang, and K. Zhao. Predicting users' continued engagement in online health communities from the quantity and quality of received support. *Journal of the Association for Information Science and Technology*, 72(6):710–722, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24436.

[203] X. Wang, K. Zhao, X. Zhou, and N. Street. Predicting User Posting Activities in Online Health Communities with Deep Learning. *ACM Transactions on Management Information Systems*, 11(3):12:1–12:15, July 2020.

[204] Y. Wang, M. Huang, x. zhu, and L. Zhao. Attention-based LSTM for Aspect-level Sentiment Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas, 2016. Association for Computational Linguistics.

[205] Y. Wang, S. Sohn, S. Liu, F. Shen, L. Wang, E. J. Atkinson, S. Amin, and H. Liu. A clinical text classification paradigm using weak supervision and deep representation. *BMC Medical Informatics and Decision Making*, 19(1):1, Jan. 2019.

[206] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49, Jan. 2018.

[207] Z. Wang, Z. Yin, and Y. A. Argyris. Detecting Medical Misinformation on Social Media Using Multimodal Deep Learning. *IEEE Journal of Biomedical and Health Informatics*,

25(6):2193–2203, June 2021. Conference Name: IEEE Journal of Biomedical and Health Informatics.

[208] X. Wei, X. Cui, N. Cheng, X. Wang, X. Zhang, S. Huang, P. Xie, J. Xu, Y. Chen, M. Zhang, Y. Jiang, and W. Han. Zero-Shot Information Extraction via Chatting with ChatGPT, Feb. 2023. arXiv:2302.10205 [cs].

[209] WHO. WHO Coronavirus Disease (COVID-19) Dashboard.

[210] M. T. Wiley, C. Jin, V. Hristidis, and K. M. Esterling. Pharmaceutical drugs chatter on Online Social Networks. *Journal of Biomedical Informatics*, 49:245–254, June 2014.

[211] W. H. World Health Organization. Advice on the use of masks in the context of COVID-19: interim guidance, 6 April 2020. 2020. Accepted: 2020-04-06T20:48:18Z Number: WHO/2019-nCov/IPC_Masks/2020.3 Publisher: World Health Organization.

[212] H. Xu, N. Zhang, and L. Zhou. Validity Concerns in Research Using Organic Data. *Journal of Management*, 46(7):1257–1274, Sept. 2020. Publisher: SAGE Publications Inc.

[213] F.-C. Yang, A. J. Lee, and S.-C. Kuo. Mining Health Social Media with Sentiment Analysis. *Journal of Medical Systems*, 40(11):236, Sept. 2016.

[214] K.-C. Yang, O. Varol, C. A. Davis, E. Ferrara, A. Flammini, and F. Menczer. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1):48–61, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbe2.115.

[215] X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, M. G. Flores, Y. Zhang, T. Magoc, C. A. Harle, G. Lipori, D. A. Mitchell, W. R. Hogan, E. A. Shenkman, J. Bian, and Y. Wu. GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records, Dec. 2022. arXiv:2203.03540 [cs].

[216] A. H. Yazdavar, M. S. Mahdavinejad, G. Bajaj, W. Romine, A. Sheth, A. H. Monadjemi, K. Thirunarayan, J. M. Meddar, A. Myers, J. Pathak, and P. Hitzler. Multimodal mental health analysis in social media. *PLOS ONE*, 15(4):e0226248, Apr. 2020. Publisher: Public Library of Science.

[217] Z. Yin, L. M. Sulieman, and B. A. Malin. A systematic literature review of machine learning in online personal health data. *Journal of the American Medical Informatics Association*, 26(6):561–576, June 2019.

[218] C. Yuan, Q. Xie, and S. Ananiadou. Zero-shot Temporal Relation Extraction with ChatGPT, Apr. 2023. arXiv:2304.05454 [cs].

[219] X. Yuan and A. T. Crooks. Examining Online Vaccination Discussion and Communities in Twitter. In *Proceedings of the 9th International Conference on Social Media and Society*, SMSociety '18, pages 197–206, New York, NY, USA, 2018. ACM. event-place: Copenhagen, Denmark.

[220] M. A. Zakkar and D. J. Lizotte. Analyzing Patient Stories on Social Media Using Text Analytics. *Journal of Healthcare Informatics Research*, 5(4):382–400, Dec. 2021.

[221] L. Zhang, M. Hall, and D. Bastola. Utilizing Twitter data for analysis of chemotherapy. *International Journal of Medical Informatics*, 120:92–100, 2018.

[222] L. Zhang, S. Wang, and B. Liu. Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4):e1253, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1253.

[223] Y. Zhang, Y. Zhang, P. Qi, C. D. Manning, and C. P. Langlotz. Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association*, 28(9):1892–1899, Sept. 2021.

[224] K. Zhao, J. Yen, G. Greer, B. Qiu, P. Mitra, and K. Portier. Finding influential users of online health communities: a new metric based on sentiment influence. *Journal of the American Medical Informatics Association: JAMIA*, 21(e2):e212–218, Oct. 2014.

[225] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision*, 130(9):2337–2348, Sept. 2022.

[226] S. Zhou, N. Wang, L. Wang, H. Liu, and R. Zhang. CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. *Journal of the American Medical Informatics Association: JAMIA*, 29(7):1208–1216, June 2022.

[227] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models, Apr. 2023. arXiv:2304.10592 [cs].

[228] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang. Can Large Language Models Transform Computational Social Science?, Apr. 2023. arXiv:2305.03514 [cs].

[229] A. Zunic, P. Corcoran, and I. Spasic. Sentiment Analysis in Health and Well-Being: Systematic Review. *JMIR Medical Informatics*, 8(1), Jan. 2020.

Appendix A

A.1 Histology used to retrieve the patient cohorts

COMPOSITE HODGKIN AND NON-HODGKIN LYMPHOMA,

HODGKIN LYMPHOMA, LYMPHOCYTE DEPLETION, NOS,

HODGKIN SARCOMA,

HODGKIN LYMPHOMA, LYMPHOCYTE-RICH,

HODGKIN LYMPHOMA, NOS,

HODGKIN GRANULOMA,

HODGKIN LYMPHOMA, LYMPHOCYTE DEPL, RETICULAR,

HODGKIN LYMPHOMA, NODULAR SCLEROSIS, GRADE 1,

HODGKIN LYMPHOMA, MIXED CELLULARITY, NOS,

HODGKIN LYMPHOMA, NODULAR SCLEROSIS, GRADE 2,

HODGKIN LYMPHOMA, NODULAR SCLEROSIS, NOS,

HODGKIN LYMPHOMA, NODULAR LYMPHOCYTE PREDOMIN,

HODGKIN LYMPHOMA, LYMPHOCYTE DEPL, DIFF FIBRO,

HODGKIN LYMPHOMA, NODUL SCLEROSIS, CELL PHASE,

MALIGNANT LYMPHOMA, LRGE B-CELL, DIFFUSE, NOS,

MALIG LYMPHOMA, LRG B-CELL, DIFF, IMMUNO, NOS,

ALK POSITIVE LARGE B-CELL LYMPHOMA,

MEDIASTINAL LARGE B-CELL LYMPHOMA,

PRIMARY EFFUSION LYMPHOMA,

INTRAVASCULAR LARGE B-CELL LYMPHOMA,

T-CELL/HISTIOCYTE RICH LARGE B-CELL LYMPHOMA,

LARGE B-CELL LYMPHOMA ARISING IN HHV8-ASSOCIATED MULTICENTRIC CASTLEMAN DISEASE,

B-CELL CHRON LYMPHOCYTIC LEUK/SMALL LYMPHOMA,

PROLYMPHOCYTIC LEUKEMIA, B-CELL TYPE,

PROLYMPHOCYTIC LEUKEMIA, T-CELL TYPE,

PROLYMPHOCYTIC LEUKEMIA, NOS,

MALIGNANT LYMPHOMA, SMALL B LYMPHOCYTIC, NOS,

PRIMARY CUTANEOUS FOLLICLE CENTRE LYMPHOMA,

FOLLICULAR LYMPHOMA, GRADE 1,

FOLLICULAR LYMPHOMA, GRADE 2,

FOLLICULAR LYMPHOMA, GRADE 3,

FOLLICULAR LYMPHOMA, NOS,

MALIGNANT LYMPHOMA, LYMPHOPLASMACYTIC,

WALDENSTROM MACROGLOBULINEMIA,

SPLENIC MARGINAL ZONE B-CELL LYMPHOMA,

MANTLE CELL LYMPHOMA,

MARGINAL ZONE B-CELL LYMPHOMA, NOS,

HAIRY CELL LEUKEMIA,

BURKITT CELL LEUKEMIA,

BURKITT LYMPHOMA, NOS,

PLASMABLASTIC LYMPHOMA,

PLASMACYTOMA, NOS,

MULTIPLE MYELOMA,

PLASMA CELL LEUKEMIA,

IMMUNOGLOBULIN DEPOSITION DISEASE,

PLASMACYTOMA, EXTRAMEDULLARY,

T-GAMMA LYMPHOPROLIFERATIVE DISEASE,

MYCOSIS FUNGOIDES,

MATURE T-CELL LYMPHOMA, NOS,

SUBCUTANEOUS PANNICULISTIC T-CELL LYMPHOMA,

HEPATOSPLENIC GAMMA-DELTA CELL LYMPHOMA,

ADULT T-CELL LEUKEMIA/LYMPHOMA (HTLV-1 POS),

PRIM CUTANEOUS CD30+ T-CELL LYMPHOPROLIF DIS (PRE-2021 CASES),

ANAPLASTIC LRG CELL LYMPH, T & NULL CELL TYPE,

INTESTINAL T-CELL LYMPHOMA,

SEZARY SYNDROME,

CUTANEOUS T-CELL LYMPHOMA, NOS,

NK/T-CELL LYMPHOMA, NASAL AND NASAL-TYPE,

PRIM CUTANEOUS CD30+ T-CELL LYMPHOPROLIF DIS,

T-CELL LARGE GRANULAR LYMPHOCYTIC LEUKEMIA,

AGGRESSIVE NK-CELL LEUKEMIA,

CHRONIC LYMPHOPROLIFERATIVE DISORDER OF NK-CELLS,

PRIMARY CUTANEOUS GAMMA-DELTA T-CELL LYMPHOMA,

T-CELL LARGE GRANULAR LYMPHOCYTIC LEUKEMIA,

LYMPHOID LEUKEMIA, NOS,

LYMPHOPROLIFERATIVE DISORDER, NOS,

MALIGNANT LYMPHOMA, NOS,

MALIGNANT LYMPHOMA, NON-HODGKIN, NOS

A.2 Codebook

Code	Definition	Example
Not relevant to personal mask-wearing		
Not relevant to facial masks	The post does not refer to facial masks, though including mask-related keywords	"The government has been masking the fact that the itself is a failure"
Not relevant to mask-wearing behavior in the COVID-19 context	Though the post is relevant to facial masks, but refers to aspects other than mask wearing, e.g., mask manufacturing, mask import and export, international news about facial masks, selling masks, etc.	"France somehow increased mask production within their country by 8 million a day. They ordered 1 billion (with a 'b') masks from China. They ordered 5 million, 15-minute-result tests. " "Amazon is deleting some third-party listings that increased surgical mask prices as the coronavirus creates a shortage"
No personal opinions		
Only sharing information and no personal opinions expressed	The post does not contain any personal opinions toward whether or not people should wear masks.	"Coronavirus: Should you wear a mask?" "EverydayHealth: Will Wearing a Face Mask Protect You From Catching the #Coronavirus?"
Personal opinion was expressed but whether it was pro- or anti-masking is not discernable.	The personal opinion expressed is ambiguous, or hard to discern whether is pro or against mask-wearing.	"Facial mask is such as heated debate now. LOL"

Table A.1: Identifying tweets that are not relevant to personal opinions toward mask-wearing.