Understanding Public Perception of Societal Concerns Using Social Media Platforms

By

Yongtai Liu

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

December 17, 2022

Nashville, Tennessee

Approved:

Bradley A. Malin, Ph.D.

Murat Kantarcioglu, Ph.D.

Tyler Derr, Ph.D.

Zhijun Yin, Ph.D.

To my parents, my wife,
and my son

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my academic advisor, Dr. Bradley A. Malin, for his inspiration, encouragement, and trust. Without his patient guidance and valuable support, I won't make it this far. I am proud to work with such a brilliant and diligent researcher.

I would also like to thank my committee members: Dr. Murat Kantarcioglu, Dr. Tyler Derr, and Dr. Zhijun Yin. Dr. Murat Kantarcioglu inspires me and gives me timely suggestions throughout my Ph.D. studies. I am grateful to Dr. Tyler Derr for his kindness and generous offer of his knowledge and expertise for my work. I must thank Dr. Zhijun Yin for his guidance and help, and for giving me encouragement to carry me through the challenges.

It is my honor to work with and learn from my excellent collaborators: Dr. Yevgeniy Vorobeychik, Dr. Ellen Wright Clayton, Dr. Abel Kho, and Dr. David Carrell. I would also like to thank my great colleagues and alumni from Health Information Privacy Laboratory for sharing their knowledge, enthusiasm, and advice.

I want to thank my friends for their help, encouragement, and the memorable moments we spent together.

Last but not least, a special thanks goes out to my family, especially my parents and wife. I am grateful to my parents for their unlimited love and support. Their cherished beliefs support me to keep moving forward. Meeting and marrying my wife was the best thing during my doctoral study. Her company and encouragement made this journey full of beauty, warmth, and happiness.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Motivation

Public perception of an object represents the way people think about it or the impression people have of it. Societal concern refers to the social events that have the potential to generate significant socio-political impact and produce social hazards [2]. Public perception of social events and societal concerns reflects social norms and values, including privacy protection and ethical awareness. For instance, studies [3, 4] have shown that the public is concerned about the possible privacy violation and familial disruption from Direct-to-Consumer Genetic Testing (DTC-GT). In a survey of 4,272 US adults [5], 33% of respondents opposed DNA testing companies sharing customers' genetic data with law enforcement.

Gaining intuition into the public's perception of societal concerns is critical for policy-makers in public decision-making [6, 7]. First, public perception profoundly impacts all aspects of social life, such as health and living habits. In a survey designed by Callaghan and colleagues [8], they found that US rural residents are significantly less likely to participate in COVID-19-related preventive health behaviors (e.g., worn a mask, avoided dining at restaurants). Second, public perception provides information needed to "address organizational and service issues" [9]. For example, opinion polls have been used to learn public views on social security legislation, people's knowledge of public affairs and government machinery [10], and various socio-political events [11]. Third, public perception reflects public needs and expectations. For instance, studies have shown that mining pubic perception in the face of natural disasters can aid in post-disaster management, health resource allocation, and government decision-making [12, 13].

Today, public perception plays an increasingly important role in public policy decision-

making and government resource allocation. In 2016, public opinion drove the "Brexit vote", which resulted in the withdrawal of the United Kingdom from the European Union. In 2019, the rising global climate movement prompted governments to take action against climate change. In 2020, about 20 million people participated in the protests in the United States against police brutality and calling for social justice and racial equality. However, researchers [6, 14] found that public perception was not well received or weighted by policymakers in the decision-making process. Gilens and colleagues [14] demonstrated that economic elites were more influential than average citizens in government policy-making. Dieckmann and colleagues [6] showed that public perception of social and cultural impacts was often not well understood and expressed during decision-making. Considering the importance of public perception in socio-political events, there is a need for better analysis and understanding of public perception of societal concerns.

Traditional studies on public perception have mainly relied upon formal surveys. However, certain flaws in the survey approach limited the insights provided by these studies. First, surveys often suffer from a limited number of respondents, which might lead to a sampling bias (e.g., undercoverage bias) [15]. Furthermore, the collected response might be shaped by question designers or hindsight bias instead of revealing the respondents' true opinion [16]. Finally, surveys are limited in their ability to shed light on the matter because they are time-consuming, and the findings (as well as the policies made upon them) can become stale in the face of the rapid evolvement of the situation.

The rise of social media provides a new research direction for public perception analysis. With the rapid development of mobile communication and the Internet, people are sharing and discussing their views, opinions, and experiences on all kinds of topics on social platforms. For example, Twitter is one of the largest social platforms in the US, with over 200 million daily active users [17]. Another notable platform is Reddit, an online content rating and discussion site where users can create different subreddits based on specific topics of interest, with over 2 billion comments posted in 2020 [18]. The large amount

of user-generated data has been relied upon to study different aspects of human life, including health and wellness [19]. For instance, Liu and Yin [20] analyzed the association between weight loss progress and Reddit users' online interactions; Klein and colleagues [21] utilized Twitter data to identify potential cases of COVID-19 in the United States. It is natural to hypothesize that this large and diverse collection of user-generated data provides opportunities to investigate public perceptions of various societal concerns.

In this dissertation, we aim to investigate how to utilize user-generated data to understand public perception of societal concerns. We select three representative social events as our research goals. In particular, we want to answer the following questions: What are the patterns and privacy risks of medical research cohort membership disclosure? Why are genetic testing consumers sharing face images and testing results online? What is the public sentiment on COVID-19-related topics, and are there any differences between urban and rural residents? These three questions cover a number of topics, including privacy, ethics, health, politics, and the economy. The third question also investigates the shift of public opinion of different population groups during the pandemic.

1.2 Research Aims

Figure 1.1 summarizes the three aims and associated natural language processing approaches of this dissertation. Our investigation revolves around two core questions: 1) *Given the societal concern of interest, what are the topics of public concern*, and 2) *what is the public sentiment about the societal concern and related topics*. We initiate our study by identifying the key topics related to the targeted societal concern. This is because the public's interest in societal concerns is typically not limited to the issue itself but includes its origin, impact, and other aspects. For instance, in the study about DTC-GT online forum, we observed that users in the forum were interested in a wide range of related topics, including ancestry composition, kinship and family finding, and genetic testing algorithms. After identifying relevant issues, we proceed to analyze public sentiment about these is-

sues. For example, in our study of COVID-19 and related topics, we inferred that US rural Twitter users had a negative sentiment towards COVID-19 prevention strategies and vaccinations.

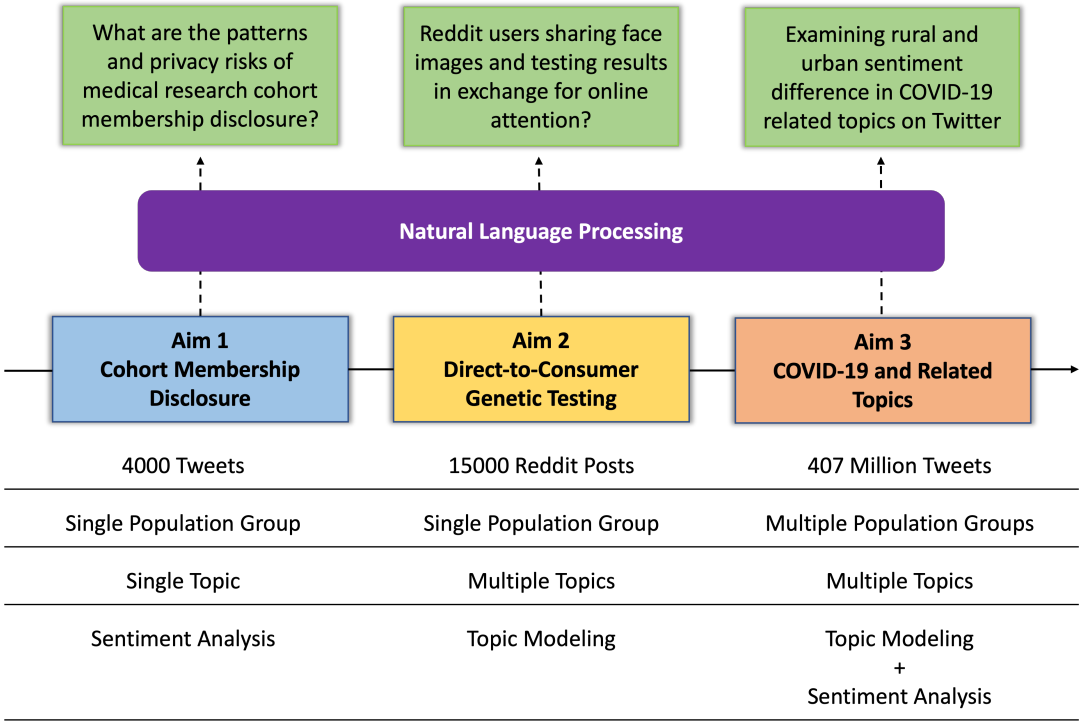

Figure 1.1: The summary of three aims investigated in this dissertation.

We unfold our investigation by solving the two core questions on three different societal concerns using data collected from two social platforms. There is a progressive relationship between the three aims regarding study size and research objective. The three aims gradually evolve from small to large, single to diverse. The first aim focuses on a single topic. In the second aim, we further study the public perception of multiple topics related to a social issue. Finally, in the third aim, we finalize the investigation by analyzing public perceptions of different population groups on multiple topics. Moreover, we believe the users of our system can readily reuse our approach in the third aim to investigate other societal concerns without additional data collection or model training. It has the potential to serve as a one-stop solution for public perception analysis on any issue of interest in a given time period.

The progressive relationship between the three aims is also reflected in the applied computational approaches. In the first aim, we relied upon online data streaming and filtering to identify the topic of interest, and analyze opinion with a rule-based sentiment analysis system. In the second aim, we utilize topic modeling techniques for topic identification and employ statistical inference to obtain statistically significant results. Finally, in the third aim, we apply a combination of clustering, text classification, word embedding, and statistical inference to study the shift of public perception of COVID-19 and related topics during the pandemic. We present detailed computational approaches, including data collection, topic modeling, word embedding modeling, and inference, when describing each individual research aim.

Aim 1: Biomedical research cohort membership disclosure on social media

The first research aim originated from a real case of membership disclosure leading to identity re-identification. Based on demographic data provided by a study cohort participant in an anonymous news interview, we successfully identified this participant in the study cohort from the cohort database. Even though this finding is by chance, we were worried about the phenomenon of online cohort membership disclosure and the potential privacy risks. This is because membership disclosure can jeopardize the privacy of the participants themselves, the reputation of the projects, sponsors, and the research enterprise.

To investigate the dangers of self-disclosure behavior, we gathered and analyzed 4,020 tweets, and uncovered over 100 tweets disclosing the individuals' memberships in over 15 programs. Through sentiment analysis, we found that 39 out of 86 (45.3%) self-disclosed users have a positive attitude towards joined research project. The terms "proud", "interest", and "love" were communicated by multiple self-disclosers. The personal information reported in the profiles of the social media users increased the risk of identification, which increases the likelihood that an attacker could link to their record in a de-identified dataset about the cohort, leading to further privacy intrusions, such as the re-identification of genomic information. A program may disclose participants' membership when they introduce

volunteers and share their stories to the public as a way to increase program influence and recruit more participants. These stories may contain personal information and sensitive health information about the volunteer.

Aim 2: Analysing association between personal information sharing and online attention received by DTC-GT consumers

Our first aim focus on membership disclosure behavior, a particular case of online self-information disclosure. Our second goal is to gain a deeper understanding of the more general phenomenon of self-disclosure. One representative research object to studying self-information disclosure is the online information sharing of DTC-GT consumers, as we observed that DTC-GT users are increasingly posting full-face images with their DTC-GT results on social platforms. Compared to membership disclosure, DTC-GT consumers shared much more detailed personal information, including genetic testing results, phenotypic traits, family history, and personal images. Investigating online posts of DTC-GT consumers also helps us gain intuition into public perception of DTC-GT, the associated benefits and concerns.

In this study, we investigated the trend in face image and testing result sharing behavior in the r/23andme subreddit to obtain insight into potential underlying motivations. Our findings show that such behavior began in September 2019 and experienced rapid growth, with over 849 face-revealing posts by early 2020. Furthermore, our study suggests that posts including a face received, statistically significant more comments and higher karma scores than other posts. Topic modeling revealed that posts that included face images were primarily about sharing and discussing ancestry composition and sharing family reunion photos with relatives discovered via DTC-GT. These findings validate our hypothesis that posting a personal image is associated with receiving more online attention, which is consistent with previous findings that people appear to be willing to give up their privacy (i.e., their personal images) in exchange for a benefit (i.e., attention from others).

Aim 3: Examining rural and urban sentiment difference in COVID-19 related

topics on Twitter

Our Aims 1 and 2 studied two specific groups: research cohort participants and genetic-testing consumers. However, to fully understand public perception of societal concerns, a topic that has wider and deeper implications for the entire population group is needed. As the COVID-19 pandemic has persisted for over two years, more than 600 million people around the world have been infected with COVID-19 with notable disparities. The COVID-19 epidemic has profoundly affected all aspects of our human life. In this respect, we investigate US Twitter users' perception of COVID-19 and related topics in the third aim. The investigation combines word embedding models with clustering strategies to identify topics closely related to COVID-19, and relied upon the similarity between topic hashtags and opinion adjectives to infer the sentiment with respect to the identified topics.

In this study, we introduced a novel approach to characterize the public's sentiment about COVID-19 and related topics. By applying topic recognition and subsequent sentiment analysis, we discovered a clear difference between US urban and rural users in their sentiment about COVID-19 prevention strategies, misinformation, politicians, and the economy. While these findings might not be representative of the sentiment of the American public more broadly, we believe that such investigations could help policymakers obtain a more comprehensive understanding of the sentiment difference between urban and rural areas on COVID-19 and related topics, so that more targeted deployment of epidemic prevention efforts can be made. Finally, we wish to highlight that our approach is not limited to COVID-19, and it can readily be extended to other topics of interest without additional data collection or model training.

1.3 Dissertation Structure

The remainder of this dissertation is organized as follows. We survey the related works in Chapter 2. The details of the aforementioned research aims are presented in Chapters 3, 4, and 5, following the convention of introduction, related work, data and method, results,

and discussion. Specifically, in Chapter 3, we investigate the behavior of membership self-disclosure and discuss the consequences. In Chapter 4, we analyze the association between face image sharing and online attention received by DTC-GT consumers. In Chapter 5, we present the solution for a multiple-group, multiple-target sentiment analysis system and discuss urban and rural sentiment difference on COVID-19 related topics. We conclude the dissertation in Chapter 6 by summarizing the main contributions and discussing limitations and future investigations. Our work has been published in AMIA [22, 23], and JMIR-Infodemology [24].

CHAPTER 2

Literature Review

As stated in chapter 1, our investigation consists of two steps: 1) topic identification and 2) sentiment inference. These two steps correspond to two fields in natural language processing: topic modeling and sentiment analysis. In this chapter, we survey the recent studies in these fields. Additionally, as we rely upon word embedding models in the third aim, we also review the applications of word embedding models related to our study.

2.1 Topic Modeling

Topic modeling refers to the process of inferring the concept (topic) of a document. For a given document, the goal of topic modeling is to generate a probability distribution of the document over k topics. The generated topic distribution can be viewed as a vector representation of the document, which often serves as an input feature for downstream tasks such as text classification. Traditional topic modeling approaches are based on TF-IDF [25] and singular value decomposition (SVD), where the TF-IDF matrix is decomposed by SVD to learn the latent topic distribution [26]. Modern topic modeling takes a "generative probabilistic approach" [27], which provides a different view for topic modeling: a document is a mixture of topics. The words are generated from the topic with probability, and the goal of topic modeling is to identify latent topics. In this subsection, we focus on probabilistic topic models.

2.1.1 Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) [28] is one of the most cited generative probabilistic topic modeling method. It treats document as a mixture of latent topics, the words are generated from the topic with probability. Both the distribution of topics in a document θ and the distribution of words in a topic ϕ follow Dirichlet-categorical distribution. The

priors α and β in these two Dirichlet functions are set to numbers much less than 1, to generate sparse words/topics distributions. These settings follow the intuition that each document has a small number of topics where each topic has few words associated with it.

Despite the success of LDA in many fields [27, 29, 30], this method still suffers from some drawbacks. First, the LDA model treats each topic independently without considering the correlation among topics. For example, to model papers in a computer science conference, publications in NLP topic would have a higher correlation with publications in the Computer Vision (CV) field than the ones in Human-Computer Interaction (HCI) field. Failing to consider such topic correlation would result in poor topic modeling quality, reflected in a high model perplexity [31, 32]. Second, LDA treats each document independently without considering the correlation between documents. Each document has several features, such as the author, the date, and the published platform (e.g., journals). The correlation among documents can be expressed by common authors, chronological order, same platform, etc. Failing to represent document correlation not only affects topic modeling quality but also adds extra effort to analyze topic modeling results: researchers need to perform a post-hoc analysis to study the association between the topic and the document feature of interest. Third, LDA has limited performance in short text topic modeling, such as social media post analysis. This is because LDA relies upon document-level word co-occurrence to infer latent topics, but social media posts are short and noisy, resulting in insufficient word co-occurrence information [33].

2.1.2 Correlated Topic Model

Blei and Lafferty [31] proposed Correlated Topic Model (CTM) to take into account topic correlation. In CTM, the Dirichlet distribution for topic prevalence is replaced by a logistic normal distribution transformed from a multivariate normal random variable [32]. If there are k topics, then the covariate matrix in the multivariate normal distribution has a size of $k \times k$, where each entry represents a topic-to-topic correlation. The logistic normal

serves as a softmax function: converts the k-sized variable sampled from multivariate normal into a vector of probabilities. Although CTM only models the correlation between two topics at most, it outperforms LDA in the quality of generated topics [31, 32].

2.1.3 Structural Topic Model

Structural Topic Model (STM) [34] improves CTM by incorporating the aforementioned document-level correlation. It assumes that both topic and word distributions are conditioned on document meta features (e.g., author, date). Suppose there are m document meta features and k topics, STM models document level information with a $m \times k$ document-topic correlation matrix. The document vector d is the multiplication of $1 \times k$ document meta vector and $m \times k$ document-topic correlation matrix [35]. Similar to CTM, distribution over topics (topic prevalence) is modeled as a logistic normal distribution, where the mean is the document vector d, and the covariate matrix is the same as in CTM. The distribution over words (topic content) is a multinomial logit of three deviation features: topic, document-level covariates, and topic-covariate interaction [34, 35]. One additional feature of STM is that it allows adding continuous meta data (such as time) as covariates to the system, enabling more flexible analysis [36]. In summary, STM allows researchers to compare the difference in topics between different document groups (author, partisan affiliation) and track the temporal trend of topics.

We relied upon LDA and STM in long document topic modeling: DTC-GT Reddit post analysis [24] and Gene-edited babies Reddit post analysis [37], respectively. Given the drawback of LDA in short document processing, we adopted a different approach in our COVID-19 Twitter analysis (aim 3): topic identification through Twitter hashtag vector clustering. The hashtag vectors are obtained from a Twitter word embedding model, which utilizes word co-occurrence information in a neighborhood instead of the whole document to learn the semantics between words.

2.2 Word Embedding

How to model words mathematically is a fundamental problem in NLP. Conventional NLP methods convert each word into a feature vector through one-hot encoding [38]. But this approach has several drawbacks: 1) one-hot encoding suffers from data sparsity; 2) new words that do not appear in the training corpus can not be well handled, and 3) the feature vector does not possess any semantic meaning. Word embedding solved these problems by representing words using a continuous vector that stores semantic and syntactic meanings. According to [39], the family of word embedding models can be divided into two groups,context-counting models and context-predicting models. In this section, we briefly review several highly cited word embedding models in these two approaches, common model evaluation methods, and applications of word embedding models.

2.2.1 Word Embedding Models

Context-Predicting Models Skip-gram model [40] trains word embedding by predicting context words around the target word. Given the target word w_t and window size c, the goal is to maximize the log probability $\sum_{-c\leq j\leq +c, j\neq 0} \log(P(w_{t+j}|w_t))$. Here, $P(w_{t+j}|w_t)$ is defined using the softmax function. However, the training speed of the skip-gram is slow since the softmax function needs to go through all the words in the vocabulary ($||V||$) to get $P(w_{t+j}|w_t)$. Mikolov and colleagues [41] then proposed negative sampling to approximate the softmax function. Negative sampling compares the training sample with only K randomly selected negative samples using logistic regression, thus the time complexity is reduced from $O(V)$ to $O(K)$.

One problem with the skip-gram model is that rare words might not be well trained, even with the help of sub-sampling the frequent words. Fasttext [42] extends on skip-gram by adding subword (or character-level) information into the model. The intuition is that character-level information can improve vector representations as "many word formations follow rules" [42]. In fasttext, a word is represented by its character n-grams. A word vector

is the sum of the vectors of its *n*-grams. Experiments showed that fasttext outperformed skip-gram in syntactic analogy tasks but not semantic analogy ones [42].

Context-Counting Model Predicting models utilize word co-occurrence information in a local context window. On the other hand, counting models assume that the statistics of word occurrences in the corpus is the core of generating good word representations. Glove [43] combines global occurrence matrix with local content window and claims that Glove outperformed skip-gram in various tasks. But Lai and colleagues [44] demonstrated that this claim could be mostly attributed to the different number of training iterations, where [43] compared a twenty-five-iterations Glove model to a one-iteration skip-gram model. In practice, Levy and colleagues [45] found that the performance difference between the two algorithms is insignificant.

Sentiment-Specific Word Embedding Word embedding models learn both syntactic and semantic information from the corpus. However, for some tasks, not both types of information are necessary. As Tang and colleagues [46] pointed out, the syntactic information in word embedding vectors can be problematic for sentiment analysis tasks. For instance, the vectors of word *good* and *bad* have high similarity due to their similar syntactic structure, but sentiment analysis task can not benefit from this similarity as the two words have opposite sentiments. Tang and colleagues [46] proposed Sentiment-Specific Word Embedding (SSWE) to focus on learning the sentiment information of words. Given the vector representation of *n* adjacent words, the training objective function aims to minimize the cross entropy loss of the predicted sentiment polarity. To acquire the sentiment-annotated training corpus, the authors collected 10 million tweets with positive and negative emoticons. Experiments showed that SSWE outperformed top system [47] in SemEval-2013 Twitter sentiment classification track.

Predicting sentiment polarity improves the performance of SSWE on sentiment-related tasks, but it also limits the generalization ability of SSWE, making it less suitable for semantic- or syntactic-related tasks. In our approach, we utilize annotated sentiment in-

formation in the inference step and keep the model training the same as in [41]. Thus our learned embedding can be applied in both sentiment analysis and semantic analysis.

2.2.2 Word Embedding Model Evaluations

Given the different choices in modeling algorithms and in hyper-parameter selection, several word embedding model evaluation methods have been proposed. Lai and colleagues [44] grouped evaluation methods based on the usage of the embedding vector: embedding as semantic properties for similarity or analogy test, embedding as features for downstream tasks, and embedding as the initialization of neural networks. The first group also referred to as "intrinsic evaluators" and the remaining two groups are referred to as "extrinsic evaluators" [48].

Intrinsic Evaluators Word similarity tasks aim to evaluate the model by comparing the correlation of word vectors with similarity scores of words judged by humans. The intuition behind it is that a well-trained model will learn the latent semantic information of words, and the semantic similarity of word pairs can be measured by the correlation of word vectors (e.g., cosine similarity). The main problem with this evaluator is that the concept of similarity is often confused with relatedness [49]. The word *Trump* and *Biden* are highly related, but it does not mean these two persons are similar.

Word analogy task was proposed by Mikolov and colleagues [50], which utilized the vector geometry to find the target word in an analogous pair. The authors introduced two types of analogy tasks: syntactic task and semantic task. Despite the fact that word analogy also suffers from the subjective problem, it serves as a good benchmark test for many applications [48].

Extrinsic Evaluators Intrinsic evaluators measure word similarity by the correlation between word vectors [51], but studies have found that this evaluation might not represent the model performance in extrinsic tasks [52]. In this regard, several extrinsic evaluators have been proposed for different downstream tasks: part-of-speech (POS) tagging, Named-

entity recognition (NER) [38], and sentiment analysis [53]. In extrinsic tasks, the word vectors serve as input for the subsequent models, which can be a shallow neural network [54], or a machine learning classifier (e.g., logistic regression).

With different embedding algorithms and various evaluators available, how to select the desired model might be a "happiness trouble". However, several studies have shown that the performance of different word embedding models is similar to each other [45, 55]. Levy and colleagues [45] demonstrated that model performance improvement is acquired through hyper-parameter tuning, instead of changing embedding algorithms. Thus, in our analysis, we chose only one word embedding algorithm, but designed two evaluation tasks to tune hyper-parameters.

2.2.3 Word Embedding Applications

Word embedding often serves as input for different NLP tasks like document clustering, classification, and named entity recognition. This type of "extrinsic usage" has been well surveyed [56]. This subsection focuses on the "intrinsic usage" usage of the word embedding model.

Stereotypes Bolukbasi and colleagues [57] revealed gender stereotypes in the word embedding model trained on Google News articles. They proposed a debiasing algorithm to remove gender stereotypes from the model by removing the gender subspace components from the original word vector. Later research showed that language model inherits human stereotype (bias) from English training corpus in race [58], gender [59] and other aspects [60]. The same phenomenon exists in different languages [61], even in textbooks [62].

Even though there are unwanted biases in the word embedding model, Garg and colleagues [55] showed that the correlation of word embedding vectors could be used to quantify historical trends and social change. They studied how gender stereotypes evolved in the US in the past 100 years. They found that women's bias in word embedding models positively correlated with women's occupation percentage in various occupations (e.g., en-

gineer, nurse, dancer). In addition, by studying the correlation between women's group vector and vectors of adjectives, they found that the women's movement in the 1960s and 1970s dramatically changed how people describe women.

Diachronic Semantic Shift Several studies investigated the temporal changes in semantic meanings using word embedding models [63]. The main difficulty in measuring semantic shift is how to compare word vectors across different models trained with different corpora. To maintain the stability of temporal word embedding models, Kim and colleagues [64] trained models in an incremental updating manner. The authors used the model at time t as the starting point for training the model at time $t+1$. Other solutions [65, 66] tried to project diachronic models to the same space, by finding a linear transformation of vectors of some "anchor" terms at different times. Kulkarni and colleagues [65] used the k-nearest neighbors of the target word as anchors for model alignment, whereas Zhang and colleagues [66] applied the transformation globally through a set of predefined semantic-stable terms. Bamler and colleagues [67] modified the word embedding objective to force the alignment during the training process to avoid post-training alignment. Eger and colleagues [68] studied the semantic change of words using time-series word embedding models. Under the assumption that most of the words are semantic-stable, Eger and colleagues measured semantic shift using a vector of similarity scores between the target word and common words among the corpus. The intuition behind this approach is that the semantic information of a target word can be obtained by comparing it with other words.

Inspired by the work of Garg and colleagues [55], we hypothesize that correlation between the target word vector and adjective vectors reveals the sentiment towards the target. Moreover, following the assumption in the work of Eger and colleagues [68], we measure the sentiment information of a target word by comparing it with a set of sentiment-annotated adjectives.

2.3 COVID-19 Sentiment Analysis on Social Platforms

In this section, we review sentiment analysis studies on COVID-19 and related topics. Depending on the subject, we divide the related research into four categories: COVID-19 general, face mask, vaccine, and others.

COVID-19 General Jelodar and colleagues [69] built a Long short-term memory (LSTM) classifier to detect the sentiment of comments in COVID-19 related subreddits. By combining topic modeling (LDA) with sentiment analysis, they studied the shift of public attitudes towards various COVID-19 related topics. The authors identified over 90 topics from 563,079 comments. A human interpretation was employed to better understand the generated topics. Chandrasekaran and colleagues [70] studied the change in public sentiment towards COVID-19 related topics using Twitter. Relying on the topic modeling approach LDA and sentiment analysis tool VADER (abbreviation for Valence Aware Dictionary for sEntiment Reasoning) [71], the authors found that the public has a negative attitude towards topics like cases update, racism, and political impact. At the same time, they observed that sentiment on the economic impact, the healthcare industry, and COVID-19 prevention and treatment shifted from negative to positive. Through LDA and VADER, Wang and colleagues [72] compared public sentiment differences towards COVID-19 between California and New York Twitter users. They found that popular topics in California and New York states are similar. At the same time, tweets posted in California had stronger negative sentiment than tweets posted in New York. Jang and colleagues [73] compared the temporal trend of sentiment towards COVID-19 in the United States and Canada via Twitter. The authors obtained 60 opinion terms through aspect-based sentiment extraction [74] and aligned the trend of sentiment with Google mobility data. Valdez and colleagues [75] studied US mental health status during COVID-19 by aligning the sentiment of COVID-19 tweets with Twitter users' timeline data. The authors applied Pruned Exact Linear Time (PELT) change-point detection [76] to identify significant changes in Twitter users' sentiment. The results suggested that the decrease in timeline sentiment may indicate decreased

social well-being. In contrast, the increase in COVID topical sentiment may be caused by a priming effect. Miao and colleagues [77] applied data augmentation technique to monitor public opinion on COVID-19 intervention measures, as data augmentation excels in various natural language processing tasks [78]. Xue and colleagues [79] applied NRC Emotion Lexicon [80] to analyze Twitter users' eight types of emotions towards COVID-19 related topics. Garcia and Berton [81] analyzed the sentiment of COVID-19 tweets in English and Portuguese via CrystalFeel [82], which combined affective lexicon features and word embedding to classify the emotion behind the sentence.

Face Mask Al-Ramahi and colleagues [83] found a strong positive correlation between the volume of tweets against face-mask mandates and the number of new COVID-19 cases. Yeung and colleagues [84] applied the LDA + VADER pipeline to study public sentiment towards mask usage. They expanded their analysis by combining sentiment analysis results with users' demographic information. They inferred users' demographics (age, gender, ethnicity) from their Twitter profile (text description and image avatar) and inferred users' political affiliation based on political candidates followed by the user. Through offline change point detection and tweets/policy news review, the authors found that Republicans had a sentiment shift after Trump's face mask tweet posted on July 20, 2020.

Vaccine Sattar and colleagues [85] analyzed Twitter users' sentiment toward COVID-19 vaccines produced by different companies. Through sentiment analysis tools TextBlob and VADER, they found that about $20\% - 25\%$ of vaccine-related tweets showed a positive sentiment, while about 10% of tweets had a negative sentiment. Their investigation of the sentiment about other safety measures had similar results. Muric and colleagues [86] collected tweets containing anti-vaccine hashtags to study the news source being used to promote the spread of anti-vaccine messages and anti-vaxxers' political leaning.

Others Dr. Jiebo Luo studied the impact of COVID-19 on social life in various aspects using social media, like hoarding behavior [87], monitoring depression [88], and college students' reaction to school closing [89]. The general steps of these studies are: first,

collect social media messages between opposing groups (i.e., hoarding vs. anti-hoarding, college students vs. general public) through hashtag searching, user profile analysis, or a classifier based on a user's Twitter timeline. Next, apply topic analysis to explore latent topics of collected tweets and use sentiment analysis (LIWC or VADER) to detect user attitudes. Finally, comparing the sentiment differences between opposing groups among latent topics.

In summary, sentiment analysis approaches adopted in the aforementioned studies can be divided into two groups: 1) rule-based analysis and 2) machine learning classification. Rule-based approaches [70, 84, 85, 90, 91] use pre-built, lexicon- and rule-based sentiment analysis applications, such as TextBlob and VADER, to directly infer sentiment in tweets. However, the rule-based approach fails to leverage the contextual information in a corpus, which varies by corpus. By contrast, machine learning approaches [69, 79, 83, 88, 92, 93] can infer implicit semantic and contextual information. Yet these methods are hindered by the need for a non-trivial amount of data annotation and training.

Aiming for an accurate and 'labor-inexpensive' sentiment analysis solution, we propose a novel approach for public sentiment analysis in our third aim. Our method analyzes sentiment on a (population) group level, which combines lexicon and semantic information to quantify public sentiment with respect to the specific population of interest.

CHAPTER 3

Analyzing Research Cohort Membership Disclosure on Twitter

In this chapter, we analyze the phenomenon of research cohort membership information disclosure on social platforms. First, we discuss the privacy risks and consequences of membership disclosure, and review existing membership inference attacks. Next, we present a novel attack that utilizes phenotypic summary statistics of a medical dataset for membership inference, and evaluate the power of the attack on four research cohort datasets. Our result suggested that it is plausible that sharing phenomic summary statistics may be accomplished with an acceptable level of privacy risk. Although the harm of the attack might be controllable, cohort participants may disclose membership information by themselves. This is because cohort participants are self-disclosing their membership and discussing their participating experience on publicly accessible social platforms, such as Twitter. Through Tweets classification and sentiment analysis, we uncovered over 100 tweets disclosing the individuals' memberships in over 15 programs, in which 71 tweets expressed a neutral or positive sentiment about participation. Our investigation showed that self-disclosure on social media could reveal participants' membership in research cohorts, and such activity might lead to the leakage of a person's identity, genomic, and other sensitive health information.

3.1 Introduction

To accelerate research and improve health care outcomes, various programs are gathering health-related information from individuals to build large cohorts [94–96]. The primary objective of these programs at the early stage is to collect a wide range of data from their participants, including genomic, phenomic (via surveys and electronic medical records), and demographic information [97, 98]. The data are then made accessible to researchers to explore hypotheses, study associations, and develop new approaches to manage one's

health [99, 100]. One common nature of these programs is that they are large and getting larger with respect to the number of participants and the size of collected data. One example of such a program is the *Personal Genome Project (PGP)*, which was launched by Harvard researchers to improve the personalization of medicine [97]. This program has collected more than 10,000 genomes of participants from a variety of countries [101]. The *100,000 Genomes Project* serves as another example, which has collected the genomes of one hundred thousand British participants to improve research on rare diseases [102]. And, to investigate how genetic predisposition and environmental exposure contribute to disease development, *UK Biobank* is now generating whole genome sequencing data on over 500,000 individuals [103].

Figure 3.1: Biomedical research cohorts.

These programs aim to make data widely available, an endeavor that is realized by sharing data with trusted researchers and, at times, with the public [95]. However, the sharing of individual-level health data raises privacy concerns. This is because participants might consent to making their genome and health data available to researchers (or to the public), but not revealing their identity, which can result in unexpected economic or reputational loss [104]. As such, the majority of large cohort programs adopt strategies to protect their participants' identity [105], for example, through the application of de-identification routines.

Yet, there are concerns about the degree to which protection can be sufficiently realized in the age of big data. This is because there are various ways in which privacy may be compromised in such systems. For instance, there have been a number of re-identification attacks designed to leverage a wide range of data types [22, 106]. In 2013, Sweeney and colleagues [107] re-identified the names of more than 40% of the PGP participants by linking demographic data (ZIP code, gender, and date of birth) of de-identified records to the voter registration lists, the attack is demonstrated in Figure 3.2.

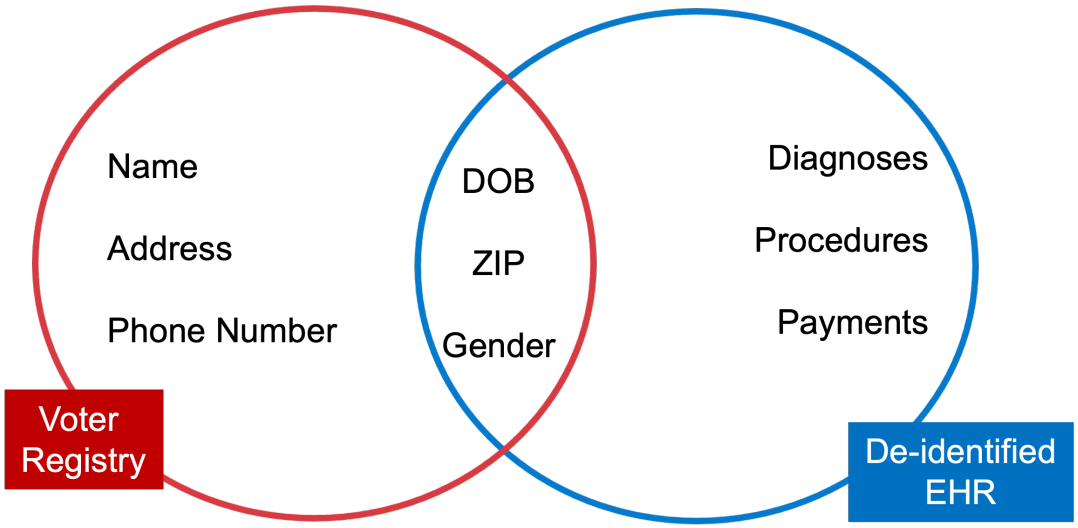

Figure 3.2: Demonstration of Sweeney's linkage attack.

Though these attacks often require a non-trivial amount of time, effort, and money to realize in a manner that would be considered detrimental to a program [108], there are several developments that are enhancing the opportunities for penetrating the privacy of individuals in such environments. The first is that participants are increasingly becoming partners in the research environment. The second, and partially an artifact of the first, is that participants are using social platforms to discuss their experiences in the research domain on a widely accessible scale [109, 110]. The third is that the research programs themselves may encourage volunteers to tell their stories publicly, with the goal of encouraging people to join the study. Revealing such information makes it evident that the social media sharer is a member of the cohort. This makes it easier for would-be attackers to identify the

sharer in the resource. This can be specifically accomplished by using the sharer's personal information that might be revealed on social media, as well as demographics that might be accessible through information brokers, to link the sharer to their record in the program's de-identified dataset [111]. The attack process is illustrated in Figure 3.3. While some individuals may feel comfortable revealing certain information about themselves (e.g., a family history of heart disease), they may not be comfortable revealing their whole genome. As such, this behavior potentially jeopardizes the privacy of the participants themselves, as well as the reputation of the project.

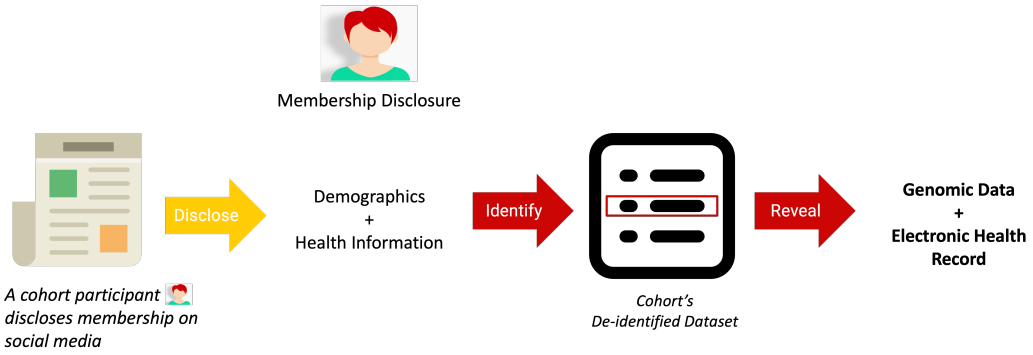

Figure 3.3: Leveraging membership disclosure for linkage attack.

To study the plausibility of an attack, we investigate the frequency of membership disclosure on social media. To do so, we selected a number of research studies from the *Database of Genotypes and Phenotypes (dbGap)* at NIH and *Wikipedia Cohort Study Category* [112]. We then set out to ascertain if any membership disclosure transpired in a popular social media platform, Twitter. To do so, we gathered over 100,000 tweets related to these cohorts and selected approximately 4,000 that contained keywords (e.g., participant, join, volunteer) indicative of potential disclosure. As will be illustrated below, we discovered membership disclosure tweets that revealed the participation of over 100 individuals. We inspected Twitter profiles for these individuals, which indicated demographics, health conditions, and occupations that might be leveraged to link to an individual's de-identified record. All of the mentioned information provides an opportunity to find the users' record in the study cohort and uncover additional information that has yet to be revealed in an

identified manner, such as the participant's genome or potentially stigmatizing health information.

This investigation also reveals several patterns. First, we show that membership-related tweets often contain certain types of words (e.g., join, participant, and volunteer). Second, over 80% membership disclosed participants have a non-negative attitude towards the program they are involved in. Sentiment analysis shows that most of these participants are happy to be a part of the cohort, which might be the incentive for some participants to reveal information about themselves. Third, longer lasting and larger cohort studies usually have more membership leakage on Twitter. We note that this is a hypothetical study only and we did not actually re-identify these individuals in the cohorts they claim to be a member of. Nonetheless, our results show that posts on social media can reveal participants' membership in research cohorts and such activity might lead to the leakage of a person's identity, genomic and other sensitive health information.

3.2 Related Work

3.2.1 Attacks Against De-identified Medical Dataset

In this section, we briefly review identity disclosure and membership detection attacks against de-deidentified medical dataset. Over time, there have been a variety of attacks perpetrated against de-identified medical and genomic data that have indicated the potential for privacy violations. Generally speaking, there are different classes of privacy attacks that have been realized, two of which are worth noting for context. The first type of attack [113] corresponds to identity disclosure. In this attack, the adversary aims to infer a person's identity (e.g., personal name) from de-identified records. One common technique employed in such an attack is linkage, where a de-identified record is related to some identified clinical records through features they commonly share, often referred to as quasi-identifiers [111, 114]. This style of attack has been applied to exploit uniqueness in genomic sequences [115], but also the combination of an individual's demographics [116],

sets of health care facilities they visited [117], laboratory test results they received [118], and diagnoses they were billed for [114]. It should be recognized, however, a linkage attack requires datasets to be disclosed in a manner that reveal individual-level data.

To protect the privacy of the participants, various programs aim to release summary statistics about certain factors about the participants in the study only. Often, these factors correspond to the rates at which certain genomic variables, such as specific alleles, were observed in the study, or in subpopulations of the study (e.g., a Caucasian or African American race). Yet, it has been shown that sharing summary statistics can still lead to privacy concerns [113]. Specifically, it was discovered that if someone has access to allele frequencies for 1) the study pool, 2) some reference population, and 3) the genomic variants of a targeted individual, then, under the right conditions, it can be predicted with some certainty if the target is the member of the pool. This type of attack, named as membership detection, is executed by comparing the similarity of the target individual's alleles to the rates at which such alleles manifest in the pool and reference (which serves as an indication of a random background of individuals from which the biased pool could have been selected), such that when the target is sufficiently similar to the pool, their membership may be disclosed. This attack, which was first posited by Homer and colleagues [119], was so surprising and successful that it prompted the NIH, as well as the Welcome Trust, to remove all genomic summary statistics from the public domain.

Several attacks has been proposed for the presence detection through genomic summary data. For instance, in the seminal work of Homer and colleagues, they measure the similarity in terms of the rate at which minor alleles are realized [119]. In the work of Gymrek and colleagues, the similarity measure relies on the distance in terms of short tandem repeats (STRs) on the Y chromosome [120]. By contrast, the attack postulated by Sankararaman and colleagues [121] relies less on a similarity measure and more on a probabilistic model, which specifically corresponds to a likelihood ratio test based on if a targeted individual is more likely to be in, than out, of a study pool. Bustamante and

colleagues [122] demonstrate a similar membership detection attack, which was applied to the Beacon platform of the Global Alliance for Genomics and Health, based on a likelihood ratio test that needs only the genomic presence data. Countermeasures for this attack, such as systematically hiding parts of the genome have been proposed[123, 124]. Wan and colleagues[108] demonstrate a way to find the optimal strategy to mitigate the privacy risk of such a membership detection attack based on game theoretic analyses using real world datasets include SPHINX and 1000Genomes.

To date, the membership attack has focused on genomic data only. While association studies have evolved to perform genome-wide scans, they have evolved to incorporate phenome- wide scans [125]. This is notable because it expands the scope of data sharing, as well as the privacy attack surface. In this respect, one must question the extent to which making summary phenome statistics available allow for presence detection attacks as well. In this chapter, we propose a new perspective of the membership attack, whereby we use a targeted individual's phenotypic information to detect their membership in a dataset. This attack diverges from previous investigations in several notable ways. First, phenotypic and genomic data are different both in terms of quality and quantity. While a person harbors on the order of 10 million basic genomic variants (e.g., single nucleotide polymorphisms or short tandem repeats), the typical structured phenomic information drawn from the clinical domain corresponds to no more than 1000 possible variants (and often orders of magnitude less). Second, we add controls to the feature space, such that we test what happens when the attacker only has partial knowledge about the targeted individual, which is a more likely occurrence in the phenotypic than genotypic scenario[116]. Third, we perform our analysis using several different datasets to create a generalizable result, which is atypical in investigations that focused on genomic data in prior studies.

3.2.2 Personal Health Information Disclosure

The personal health information that has been disclosed on social media has been leveraged to study health-related behaviors [19, 109, 126]. In spite of the great potential research value, there still exist many concerns regarding the sharing of personal health status or negative health risk behaviors in online environments [127]. For example, Morgan and colleagues [128] showed that one-third of investigated college students reported having posted a picture depicting substance use on social media platforms. Sharing such information will not only trigger privacy concerns about the disclosers themselves (e.g., damage to reputation), but may have the potential to influence other people's behavior. For instance, it was observed that discussions about prescription abuse over Twitter may aggravate substance abuse [129, 130].

Additionally, it should be noted that people share their own information as well as that of other people in online environments. It has been shown that individuals disclose information about a wide range of acquaintances, ranging from family members to friends to high profile persons in the media [19, 131]. For example, Christofides and colleagues [132] illustrated how undergraduate Facebook users posted personal information (e.g., dates of birth and email addresses) in their profiles, but also shared photos of their friends performing potentially sensitive acts (e.g., drinking alcohol at parties).

Our work differs from the aforementioned studies in that we focus on the privacy issues regarding the membership of participants in biomedical research programs on social media. Specifically, we study self-disclosures made by the program participants themselves, as well as investigate the disclosures made by the organizations who own and have responsibility to protect the participants' data. In doing so, our research contributes to the health information privacy field by highlighting a new type of privacy risk: the cohort membership leakage through social media.

3.3 Membership Detection using Summary Statistics

In this section, we present the membership detection attack through phenotypic summary statistics. We use a targeted individual's phenotype information and statistical summary data about a given dataset to detect if the targeted individual is in the dataset or not. We collect data from SPHINX, Vanderbilt University Medical Center (VUMC), Northwestern University (NW) and Kaiser Permanente Washington Health Research Institute (KPW) to perform this analysis. Specifically, we use SPHINX as the pool dataset, and the other three datasets as references. Based on these datasets, we performed a systematic series of controlled experiments to assess the risk of phenotypic membership detection. We show that this attack is possible and can achieve a relatively accurate result (over 80% precision and recall) under certain conditions. However, at the same time, we find that the strength of this attack can be controlled and is likely mitigated by natural phenomena. In particular, this is because the attack is dependent upon the adversary's knowledge about the target and reference dataset. And, as we illustrate, the attack weakens as we limit the amount of phenotypic information available to the adversary, as well as when the size of the reference dataset grows. A particularly notable finding is that when the number of phenotypic features available to the attacker is no greater than 6 (a plausible number of features available to an adversary[116]), the recall and precision of this attack is below 13%.

3.3.1 De-identified Phenotypic Datasets

eMERGE-PGx dataset The eMERGE-Pharmacogenetics (PGx) [133] dataset contains the genetic sequencing and phenotype data (demographics, medications, and ICD-9, PheWAS codes and CPT codes) of 8,173 subjects. These subjects are from 9 different eMERGE-PGx sites: Children's Hospital of Philadelphia (CHOP), Cincinnati Children's Hospital Medical Center (CCHMC), Geisinger Health System (GHS), Kaiser Permanente Washington with University of Washington (KPW), Marshfield Clinic, Mayo Clinic, the Icahn School of Medicine at Mount Sinai, Northwestern University (NW), and Vander-

bilt University Medical Center (VUMC). In particular, this dataset provides the individual records of each subject. In each record, a list of phenome-wide association study (PheWAS) codes is generated from the subject's phenotypic information [125]. A complete PheWAS code has 5 digits including the root code (i.e., the first 3 digits) and child code (i.e., the last 2 digits). Some of the PheWAS codes in the dataset only contained the root code, while others contained the root and child codes. For consistency, we rolled up every 5 digit code to its 3 digit form. For example, after the roll-up, a subject with PheWAS codes 008.51, 008.52, and 008.60 will be represented by a single code: 008. A total number of 579 PheWAS root codes appeared in the dataset at least once. The prevalence of each 3-digit PheWAS code was computed based on the individual level records. Here, prevalence is defined as the proportion of patients that were assigned the code. After preprocessing, the data consisted of the prevalence of 579 PheWAS codes.

VUMC SD Dataset The VUMC Synthetic Derivative (SD) dataset contains de-identified individual level records of over 2 million subjects [134]. These subjects are the set of patients that received any medical care at VUMC. 11.0% of the participants in eMERGE-PGx are sampled from the VUMC. For each subject, this dataset contains a list of ICD-9 codes derived from the electronic medical record. The SD is a set of records that is no longer linked to the identified medical record from which it is derived and has been altered to the point it no longer closely resembles the original record. We map each ICD9 codes to a 3 digit PheWAS code and compute the prevalence of each PheWAS code in the VUMC SD population. After preprocessing, we obtained the prevalence of 564 PheWAS codes and the individual-level data of 2,155,348 subjects.

KPW and NW Datasets The KPW and NW datasets contain the aggregated statistics of ICD-9 codes of 2,446,230 patients that received care from Kaiser Permanente Washington (KPW) and 1,602,402 patients Northwestern University Memorial Hospital (NW). 12.1% and 8.9% of participants in eMERGE-PGx are sampled from KPW and NW respectively. For this study, we rely on the counts of patients who received each ICD-9 code.

Based on the ICD-9 PheWAS mapping, we compute the count of each 3-digit PheWAS code as the sum of the count of each ICD-9 code that is mapped to the PheWAS code. This preprocessing provided the prevalence of 452 PheWAS codes for subjects from KPW, as well as NW. Table 3.7 provides a summary of the datasets.

3.3.2 Membership Detection based on Likelihood Ratio (LR) test

Based on this data, we now provide intuition into the attack. Given the set of PheWAS codes a subject has been assigned, along with the prevalence of PheWAS codes in the pool and the reference, we can make a prediction about if the target is in the pool (or reference) based on a likelihood ratio (LR) test. Formally, we represent each individual as a set of PheWAS codes in a binary vector $x = (x_0, x_1, \ldots, x_n)$, where $x_i = 1$ if the subject has the i^{th} PheWAS code and 0 otherwise. We then compute the LR score as:

$$\ln \frac{P(x|pool)}{P(x|ref)} = \ln P(x|pool) - \ln P(x|ref) \tag{3.1}$$

where $P(x|pool)$ and $P(x|ref)$ is the probability that a subject x is in the *pool* and $reference$, respectively, given their PheWAS codes prevalences.

We represent the prevalence of the i^{th} PheWAS code in the pool and reference as $P(pool,i)$ and $P(ref,i)$, respectively. We can derive $P(x|pool)$ and $P(x|ref)$, where:

$$P(x|pool) = \prod_i P(pool,i)^{x_i}(1 - P(pool,i))^{1-x_i} \tag{3.2}$$

$$P(x|ref) = \prod_i P(ref,i)^{x_i}(1 - P(ref,i))^{1-x_i} \tag{3.3}$$

And, we can represent the LR score in terms of $P(pool,i)$ and $P(ref,i)$:

$$\begin{aligned}\ln \frac{P(x|pool)}{P(x|ref)} = \sum_i x_i \ln P(pool,i) + (1-x_i)\ln(1-P(pool,i)) \\ - x_i \ln P(ref,i) - (1-x_i)\ln(1-P(ref,i))\end{aligned} \tag{3.4}$$

Now, given a predefined threshold θ, we predict that the subject is in the pool if their LR score is above the threshold. Otherwise, we predict that the individual is in the reference.

Performance measures We use standard measures to measure the performance of the attack. Recall is the fraction of correctly predicted pool subjects over the size of the pool. It measures the completeness of our prediction. Precision is the fraction of correctly predicted pool subjects over all subjects we predicted as in the pool. It measures the relevance of our prediction. Accuracy is the fraction of all the correctly predicted subjects over all subjects in the pool and reference. It measures the prediction accuracy for both the pool and the reference. In this attack, since we care more about the subjects' presence in the pool, but not the reference, accuracy is not as important of a measure as others in practice, but we report it for completeness. The F_1 score is the harmonic average of the precision and recall and is the primary privacy measure employed by our experiments. Let TP be the number of correctly predicted pool subjects, FP be the number of subjects predicted in the pool by mistake, TN be the number of correctly predicted reference subjects, P be the size of pool and N be the size of reference shown in, then $recall = \frac{TP}{P}$, $precision = \frac{TP}{TP+FP}$, $F_1\ score = 2\ \frac{precision \times recall}{precision+recall}$ and $accuracy = \frac{TP+TN}{P+N}$.

3.3.3 Experiments and Results

To perform our experiments, we set the eMERGE-PGx dataset as the pool and use VUMC SD, KPW and NW as the reference populations. We first investigated the extent to which a targeted individual can be detected in the eMERGE-PGx pool of 8,173 individuals. To perform this analysis, we randomly sampled a subset of subjects from the VUMC SD to form the reference set. We simulated scenarios with three different ratios of the pool and reference sample size: 1:1, 1:25, 1:250. As such, the reference set sizes for each scenario is 8,173, 204,325, and 2,155,348 (all subjects in VUMC SD). The details of the experimental setup are reported in Table 3.1.

In each setting, for each subject in the union of the pool and the reference, we compute

Table 3.1: Design of the experiments that assess the influence of the prior probability of a targeted individual being in the pool.

Scenario	Pool Size	Reference Size	Proportion (pool:reference)
a	8,173	8,173	1:1
b	8,173	204,325	1:25
c	8,173	2,155,348 (all subjects)	1:250

the LR test score. We make one of two claims based on the score: i) the subject is in the pool or ii) the subject is in the reference. We say that a false positive occurs when we claim the subject is in the pool, but they were really in the reference. We repeated the experiment 10 times, each time randomly sampling a portion of records from VUMC SD. The average precision, recall and accuracy of the 10 runs at 20 different threshold levels are shown in Figure 3.4. In Figure 3.4, the pool:reference sample proportion decreases from left to right. Recall and Precision are shown in the first row and Accuracy and F_1 are shown in the second row.

For every threshold θ, we compute the variance of recall and precision at 10 runs, and report the average variance of recall and precision. For Figure 3.4.a, the average variance of recall is 4×10^{-6}, the average variance of precision is 2×10^{-5}. For Figure 3.4.b, the average variance of recall is 1×10^{-7}, the average variance of precision is 1×10^{-3}. These results imply that average performance of our attack is highly stable in the context of these experiments.

There are several notable findings that are illustrated in Figure 3.4. First, it can be seen that the most successful attack setting is realized when the pool and reference are of equal size, as shown in Figure 3.4.a. In this case, the recall, precision, and F_1 score are all maximized when the threshold is $\theta = -20$ (recall = 0.839; precision = 0.828; F_1 = 0.833). This means that, at this threshold, we correctly detected the presence of more than 80% of the subjects with a false positive rate of approximately 18%.

Second, as shown in Figures 3.4.a and 3.4.b, it can be seen that the precision changes

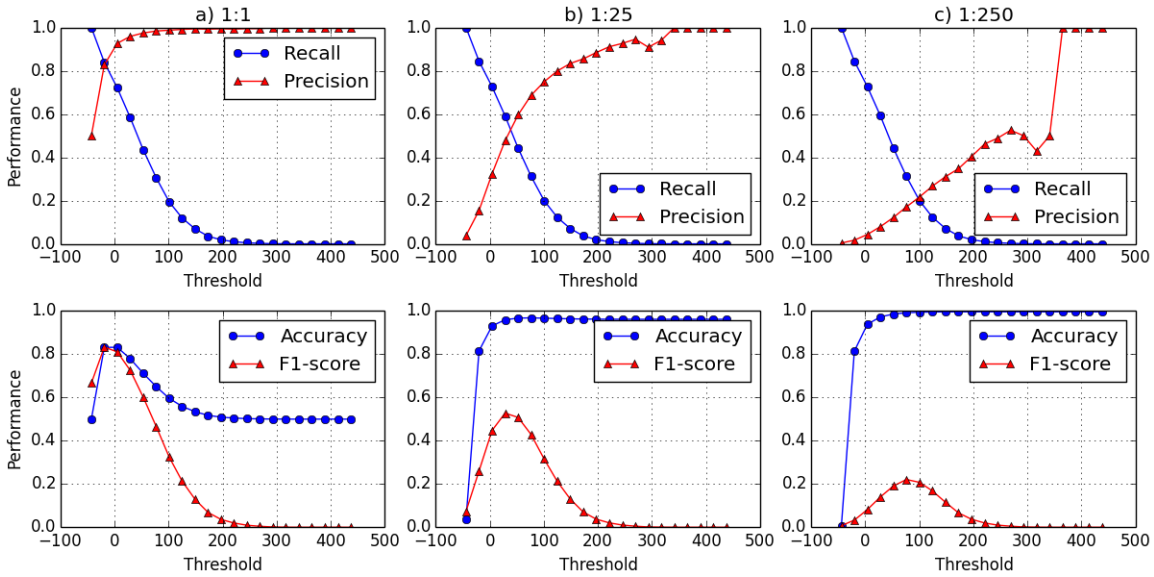

Figure 3.4: The performance of the phenotypic presence detection attack as a function of threshold θ with different pool:reference sample proportion.

significantly with changes in the pool to reference proportion. In particular, as the size of the reference grows, the initial precision value drops from 0.5 to 0.04 and eventually falls to 0.004. This is due to the change in the prior probability that a target is in the pool. When the attacker chooses the lowest threshold ($\theta \approx -50$), they will predict every subject as a member of the pool because every record has a score that is greater or equal to the lowest score. It can be seen that, as the ratio changes from 1:1, 1:25 to 1:250, the prior probability of being in the pool decreases and, as a result, the attack performance at the lowest threshold worsens. Moreover, we can see that the shape of the precision rate (as a function of the proportion) changes as well. Specifically, it shifts from a concave to a convex function. This happens because, as we increase the size of reference, more subjects in the reference appear to be similar to those in the pool. As the number of false positives grows, the precision decreases. Eventually, when the threshold is raised to a sufficiently high level, the precision becomes 1, which occurs at the point when only a few true positive samples remain.

Third, the F_1 score decreases as we increase the size of reference. This is expected

because the precision is decreasing. However, it should be recognized that the decrease in the precision and the increase in the size of the reference population are not changing at the same rate. As we increase the population size from 8,000 to 2.2 million people, the highest F_1 score drops by 75% of its initial value. We further recognize that as the size of reference increases, the threshold for the best overall performance (i.e., the highest F_1 score) increases as well. This is an artifact of the decrease in precision. This observation suggests that, when the size of the reference population is large, the adversary should rely upon a larger threshold value.

Fourth, as can be seen in Figure 3.4.b and Figure 3.4.c, the precision has a small gap at θ around 300. This happened because of the large size of the reference population (i.e., 2.2 million people). In this case, the probability that it contains several subjects that look like they came from the pool is relatively high. As a result, these subjects remain in the predicted pool until θ is sufficiently large, which occurs when θ is approximately 300. This probability reduces as the reference population becomes smaller. By comparing Figures 1.b and 1.c, it can be seen that the red line (which corresponds to the precision) in 1.b is much smoother than the one in 1.c.

In summary, the findings from this experiment indicate that, as we increase the size of the reference population, the power of the attack decreases. This is primarily due to the fact that the prior probability of a target being in the pool will decrease. Notably, this finding is in alignment with the findings of Heatherly and colleagues [135], which indicated that a larger population can offer better privacy protection than a smaller one.

eMERGE-PGx vs KPW and eMERGE-PGx vs NW Next, we conducted a similar set of experiments with the three reference datasets based on summary statistics: eMERGE-PGx vs. KPW, eMERGE-PGx vs. NW, and eMERGE-PGx vs. VUMC SD. In these experiments, we set eMERGE-PGx as the pool. For the purpose of generalizability, we randomly generate a set of subjects, with PheWAS codes. We represent these as $x = (x_1, x_2, \ldots, x_n)$, where $x_i = 1$ is the i^{th} PheWAS code in the record. We assume that each x_i is independently

randomly distributed and $P(x_i = 1)$ is the prevalence of the PheWAS code in the dataset. For each dataset, we generate 204,325 subjects, such that the proportion between the pool and reference is 1:25.

We compute the LR test score for the union of all the subjects in eMERGE-PGx and the simulated datasets. To mitigate the impact of the randomness, we generated 10 random datasets of a given size for each reference. The average precision, recall, accuracy and F_1 score are shown for the VUMC SD, KPW and NW reference datasets in Figures 3.5.a, 3.5.b, and 3.5.c, respectively. All of the variances for the recall and precision for three experiments are smaller than 1×10^{-6}. Again, this result implies that the average performance from the experiments is likely to be stable.

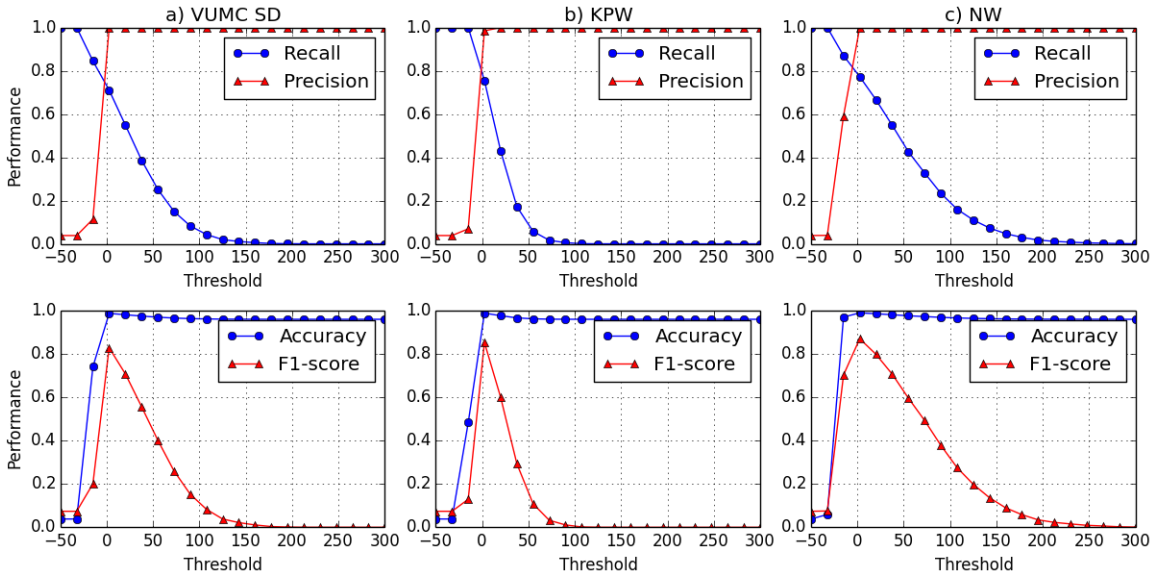

Figure 3.5: The performance of the attack as a function of the threshold θ with three different reference populations.

It can be seen that NW has the highest F_1 score (0.87), followed by KPW (0.85), and then VUMC SD (0.83). The difference in the three highest F_1 scores is not that large. However, we note that this might be an artifact of using simulated data in these experiments.

eMERGE-PGx vs VUMC SD using a Subset of PheWAS Codes The previous experiments are based on the assumption that the adversary knows the presence and absence of all the PheWAS codes for all the subjects. Thus, in this experiment, we assessed the

performance of the LR test in a scenario when the adversary's knowledge is limited to only a subset of the PheWAS codes for a targeted individual. This is a more realistic scenario because it is unlikely that an attacker will know everything about everyone[116]. In this experiment, we use eMERGE-PGx as the pool and all subjects from the VUMC SD as the reference. We conducted two sets of experiments. The first experiment relies on 5 PheWAS codes, the second experiment relies on 55 PheWAS codes. We compare these results with the earlier experiment that permitted the adversary to leverage all 557 PheWAS codes.

We performed 10 runs for each experiment. In each run, we used a set of randomly selected PheWAS code to compute the LR test score. The average results for the runs are depicted in Figure 3.6, where the number of PheWAS codes used in each experiment decreases from left to right. For Figure 3.6.b, where the adversary has 55 PheWAS codes, the average variance of recall is 7×10^{-4}, and the average variance of precision is 0.019 . For Figure 3.6.c, where the adversary has 5 PheWAS codes, the average variance of recall is 0.018, and the average variance of precision is 0.026. It can be seen that as the number of features decreases, the range of the LR score decreases as well. Specifically, the range for the threshold shrinks from [-100, 500] to [-5, 15].

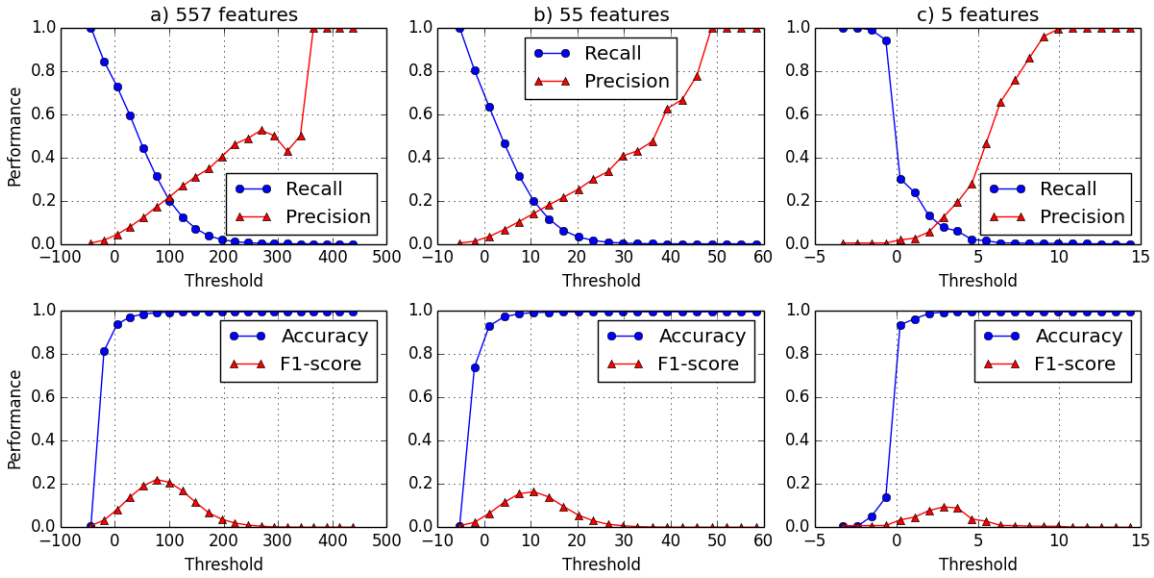

Figure 3.6: The performance of the attack as a function of threshold θ and the number of phenotypic features available to the adversary.

These results indicate that, as the number of features decreases, the precision, recall and the F_1 also decrease. When the adversary has 55 PheWAS codes, they can achieve a 19.8% recall and 14.1% precision, with a maximum F_1 score of 16.5%. However, when the adversary only has access to 5 features, the maximum F_1 score is 9.4%, with 7.6% recall and 12.4% precision. This suggests that when the attacker has limited knowledge about the targeted individual, detecting their membership in the pool becomes much more difficult. While 10% predictive power may seem large, this suggests that the adversary's success is, on average, similar to the risk tolerance often recommended by federal agencies and other various policies in practice[116]. This is notable because it further suggests that amendment to the summary statistics may not be necessary to make the data publicly accessible.

3.4 Membership Disclosure on Twitter

In the previous section, we evaluated the power of membership detection attack through summary statistics and showed that the risk is acceptable under appropriate protection. However, the activities of participants and research program sponsors, particularly on social media, might reveal an individual's membership in a study, making it easier to recognize participants' records and uncover the information they have yet to disclose. This behavior can jeopardize the privacy of the participants themselves, the reputation of the projects, sponsors, and the research enterprise. To investigate the dangers of self-disclosure behavior, we gathered and analyzed 4,020 tweets, and uncovered over 100 tweets disclosing the individuals' memberships in over 15 programs.

3.4.1 Study Cohorts Selection

We selected study cohorts from the database of *Genotypes and Phenotypes (dbGap)* and *Wikipedia cohort studies category* [112]. dbGaP was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in humans. It contains 483 biomedical research studies. By contrast, Wikipedia

provided a convenient list of long lasting cohort studies, such as the *1970 British Cohort Study* [136]. To make our investigation more general, we chose as many different types of studies as possible. The selected cohorts are diverse in three aspects: objective, time, and population.

Objective. We selected cohorts to focus both on a specific disease, such as *Type 1 Diabetes Genetics Consortium* [137], as well as a particular demographic, such as the *Nurses' Health Study* [138] or gender with in the *Million Women Study* [139].

Time. Cohort studies are not a new phenomenon. Some of the cohorts considered have a long history. For instance, the *Framingham Heart Study* [140] began in 1948. Still, some of the studies are relatively new, for example, the *Qatar BioBank* [141] was launched in 2012. Additionally, we selected studies to have a wide range in duration. Certain longitudinal studies have lasted for decades, while some achieved their objectives in a short period and thus were quite limited in length.

Population. The selected cohorts have a varying number of participants. There are multiple cohorts with relatively small sizes, such as the *International HapMap Project* [142], which collected human genomes from 1,000 participants. By contrast, several cohorts contain hundreds of thousands of participants, such as the *100,000 Genomes Project*.

3.4.2 Data Collection and Data Analysis

We partition our search procedure into two steps. Here we provide a high-level overview of the process. The first step is data collection. In this step, we find all of the possible tweets related to the selected cohorts. The second step is data filtering. In this step, we choose a portion of the tweets from step one and manually review these tweets to find the participants of studies. We then perform sentiment and frequency analysis on the tweets that disclose membership in a biomedical research study. The workflow is summarized in Figure 3.7.

Data Collection. To collect tweets related to the selected cohorts, we use the names (and abbreviations) of the 77 studies as search keywords and collect all related tweets with

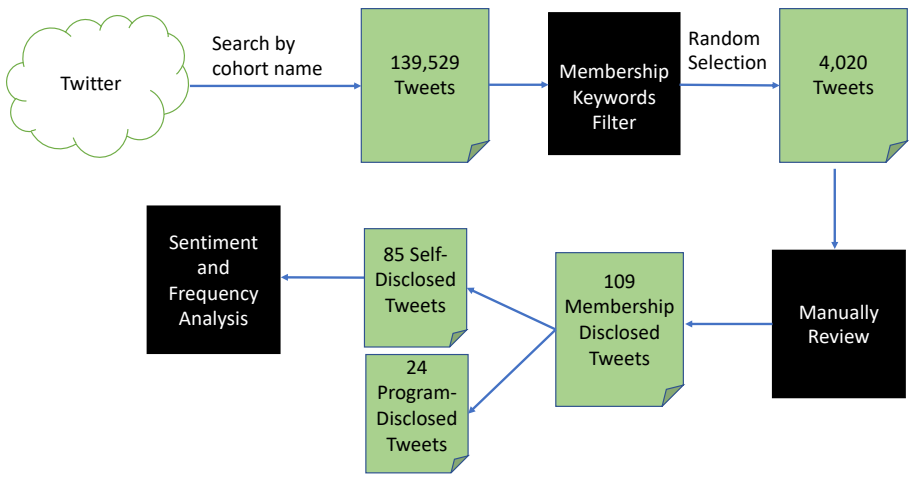

Figure 3.7: The framework for research cohort membership discovery.

a python crawler. By doing so, we obtained 139,529 tweets. Manually reviewing all of the tweets to find those revealing an individual's participation information would be quite time consuming and error-prone. Since this is a pilot study, and our goal is to demonstrate the possibility of membership disclosure instead of finding all such tweets, we narrowed the scope of our search based on our knowledge to a portion of the tweets that are most likely to contain information about membership disclosures. When we manually reviewed some of the collected tweets, we found that most of the self-disclosed tweets exhibited the following pattern: "I joined xxx research project today!", "I am a participant of the xxx program." or "Now I became a volunteer of the xxx study."

Data Filtering. We filtered the tweets with the following keywords: *participant*, *participate*, *join*, and *volunteer*, and discarded the remainder of the tweets. It should be recognized that this search method does not guarantee completeness. We lose tweets about disclosure that lack such search terms. For example, "I sent my test sample to xxx project today." is not caught by the filter. This step yielded 12,698 tweets. For most of the projects, there are fewer than 500 tweets with the keywords of interest. Thus, we manually reviewed all of these tweets to find those that reveal membership disclosure. For cohorts with more than 500 tweets, we randomly select 500 for manual review. Details about the number of

tweets collected for each cohort are provided in Table 3.2. For brevity, we depict the top 50 cohorts that returned the most tweets. Information on all 77 cohorts is available on Github[1].

Sentiment and Frequency Analysis. The previous step yielded 4,020 tweets. We manually reviewed these tweets, and labeled the tweets containing membership disclosure information. We performed sentiment and frequency analysis on the target tweets posted by project participants. We first removed all the links, hashtags and @ characters from the tweets. We then fed the preprocessed tweets into *TextBlob* (version 0.15.3) for sentiment analysis. *TextBlob* is a python package for natural language processing (NLP). For each tweet, *TextBlob* generates a sentiment score in the range from $[-1, +1]$, where -1 means extremely negative and +1 stands for extremely positive. Next, we partitioned the tweets into words through a process of normalization and tokenization (which partitions a tweet into a set of words), lemmatized (which transforms a word from its original form to its base form; e.g., walks becomes walk) all the words using python NLP package *nltk* (version 3.3). For the lemmatized words, we removed stop words (e.g. i, ia, in ,the). Since we used cohort names to collect all the tweets, we also dropped all of the words in cohort names, such as "study", "project", "health" and "genome". We then counted the frequency for the remaining words.

3.4.3 Results

Table 3.2 reports the number of tweets collected, filtered and reviewed for 77 selected cohorts. Each of the first six cohorts in Table 3.2 has more than 10,000 related tweets, which in total accounts for 70% of the total collected tweets. All of the cohorts in the top 25% have over 1,000 tweets. The number of tweets collected from these 19 cohorts accounts for 91.8% of all the tweets. There are 26 cohorts with fewer than 100 related tweets. The distribution of tweets filtered by the selected keywords is roughly the same as the distribution of the total collected tweets. The set of cohorts in the top 6 occupied 87.5% of the filtered tweets and the top 19 cohorts generated 97.6% of filtered tweets. In general,

[1] https://github.com/yongtai123/Biomedical-Research-Cohort-Membership-Disclosure

Table 3.2: Number of tweets collected, filtered, and reviewed for 77 cohorts.

	Study Cohort	All Tweets	Tweets Filtered	Tweets Reviewed
1	UK Biobank	24,056	4,265	500
2	100000 Genomes Prjoect (Genomics England)	22,217	1,735	500
3	UK 10K	14,999	515	500
4	LifeLines	14,600	74	74
5	All of Us Research Program	12,263	4,163	500
6	National Children Study	10,640	357	357
7	Human Longevity	4,478	45	45
8	Qatar Biobank	3,585	296	296
9	Australian Longitudinal Study on Women's Health	3,555	136	136
10	Research Program on Genes, Environment and Health	2,769	6	6
11	Personal Genome Project	2,655	436	436
12	Raine Study	2,538	72	72
13	Generation Scotland	2,283	106	106
14	Coronary Artery Risk Development in Young Adults Study	1,616	11	11
15	Nun Study	1,275	19	19
16	Millennium Cohort Study	1,195	42	42
17	Million Women Study	1,182	17	17
18	Socio-Economic Panel	1,152	64	64
19	Avon Longitudinal Study of Parents and Children	1,005	43	43
20	Young Lives	919	36	36
21	LifeGene	850	3	3
22	Seven Countries Study	833	3	3
23	Atherosclerosis Risk in Communities	708	2	2
24	English Longitudinal Study of Ageing	673	6	6
25	Black Women's Health Study	619	22	22
26	International Cancer Genome Consortium	601	20	20
27	1970 British Cohort Study (BCS70)	600	26	26
28	Whitehall Study	575	10	10
29	Nurses' Health Study	565	29	29
30	Alameda County Study	353	1	1
31	Seattle 500 Study	339	2	2
32	National Child Development Study	299	26	26
33	Framingham Heart Study	290	14	14
34	Religious Orders Study	249	17	17
35	The Irish Longitudinal Study on Ageing	216	7	7
36	Women's Interagency HIV Study	184	3	3
37	Adventist Health Studies	166	1	1
38	Study of Mathematically Precocious Youth	160	4	4
39	Newcastle 85+ Study	148	7	7
40	Great Smoky Mountains Study	129	2	2
41	International Rare Diseases Research Consortium	128	4	4
42	UK Households Longitudinal Study	126	1	1
43	Multicenter AIDS Cohort Study	119	3	3
44	National Survey of Health & Development	116	2	2
45	British Birth Cohort Studies	113	0	0
46	BioBank Japan	103	0	0
47	MalariaGEN	103	5	5
48	Taiwan Biobank	102	2	2
49	COSMOS Cohort Study	100	0	0
50	Normative Aging Study	100	1	1
	...	...	...	...
	Summary	**139,529**	**12,698**	**4,020**

the research programs with larger volume and longer time span have to had more tweets. In particular, programs involving government support often fall into this category.

Table 3.3: A summary of the cohort and membership coverage from tweets discovered to reveal participation.

	Study Cohort	Tweets Reviewed	Tweets Disclosed	Self-Disclosed Tweets	Disclosed Tweets	Disclosed Individuals
1	Personal Genome Project	436	26	26	0	26
2	100,000 Genomes Project (Genomics England)	500	16	9	7	18
3	Black Women's Health Study	22	12	12	0	12
4	Raine Study	72	11	0	11	14
5	UK Biobank	500	10	10	0	10
6	All of Us Research Program	500	10	10	0	10
7	Qatar Biobank	296	5	4	1	4
8	Nurses' Health Study	29	4	4	0	4
9	Australian Longitudinal Study on Women's Health	136	3	3	0	3
10	1970 British Cohort Study (BCS70)	26	3	3	0	3
11	Framingham Heart Study	14	2	0	2	2
12	Millennium Cohort Study	42	2	2	0	2
13	Million Women study	17	2	2	0	2
14	National Child Development Study	26	2	1	1	3
15	Human Longevity	45	1	0	1	1
	Summary		**109**	**86**	**23**	**114**

Among the 4,020 selected tweets, we found 109 that communicated membership disclosure. The results of this investigation are shown in Table 3.3. These tweets come from 15 of the cohorts (19.5%). They reveal the membership of more than 115 participants. We present some examples of disclosure tweets in Table 3.4. We replaced the person and cohort names with xxx and rewrite the sentences to mitigate the risk of revealing the program and participants. In Table 3.4, 86 of these tweets (78.9%) were posted by cohort participants. In these cases, participants' leaked either their own or their friends' membership information when they talked about their experience with some cohort study. This discovery confirms the findings of [143] and [131], where it was observed that individual's self-disclosure on social media may reveal other people's sensitive information. The remaining 23 tweets (21.1%) come from the program's official account or researcher/organizer of the study. In these cases, the participants' information was revealed because the program shared a volunteer's story.

We discuss self-disclosed and program-disclosed tweets separately in the following sec-

Table 3.4: Examples of membership disclosure tweets.

Type	Tweet
Self-Disclosed	1. Proud to be a participant in this: https://url/abcd 2. I like how xxx program never forget my birthday. Thanks @xx 3. I joined xxx project because ..., I won't never share anyone else's DNA. 4. I am both a researcher and a participant of the xxx project.
Program Disclosed	1. It's great to see Mr.xxx and his parents sharing their story about receiving a test result from the xxx research https://url/abcd 2. In this video, meet participant Ms.xxx and her father, xxx, who talked about why taking part is important to them https://url/abcd 3. It's awesome that @xxx continue to contribute to the Program! 4. XXX, who has heart disease, talks about her participation in xxx study .

tions.

Self-disclosed tweets. Self-disclosure tweets refer to the tweets posted by cohort participants. These tweets usually have a similar style, such as "I joined/participated in the xxx study" or "I am a participant/volunteer of the xxx program." Some users wrote an additional sentence to explain why they joined the program or how they feel about it. An analysis of the sentiment of self-disclosed tweets revealed that 71 of the 86 users (82.5%) have a neutral or positive attitude about their participation while 39 of the tweets (45.3%) have a sentiment score greater than 0. Such a positive attitude shows that most self-disclosed volunteers are happy with the program they participate in and their disclosures on social media express their support or compliment for the program rather than criticism. Words like *proud* and *love* often appears in these tweets. Table 3.5 provides the frequency of the 26 most common words.

At the same time, a small portion of the tweets suggests a negative emotion. For example: "I've been a participant for two years, but have not had any feedback." The distribution of the sentiment score is shown in Figure 3.8. Self-disclosure tweets usually only reveal the user's membership; however, at times they may involve their family or close friends. In such cases, one or more of the users' family members may have a rare disease (e.g., a child who experiences a congenital heart attack) and they joined the research project together to

Table 3.5: The most frequent words in 86 self-disclosure tweets.

word	count	word	count	word	count
participant	26	invite	6	would	4
participate	23	love	5	well	4
join	20	since	5	remember	4
get	9	data	5	today	4
proud	7	one	5	great	4
interest	6	share	5	learn	4
years	6	look	4	think	4
volunteer	6	member	4		

find out why and how to treat it.

Program-disclosed tweets. At times, the programs post about volunteers' participation experiences on social media as a way to promote the program and attract the public to join. Most of these tweets reveal a volunteer's membership often with health information, along with a link to, or a video about, the volunteer's story. Volunteers talk about why they joined the program, as well as what they gained from entering the program. This approach may be useful in attracting people to join the program, but this activity also increases the risk of the volunteer to re-identification.

Disclosure tweets are more likely to be associated with larger cohorts. As shown in Table 3.6, the cohorts with membership disclosure tweets cover more than 10,000 participants. The studies that began more recently tend to have more members active on the Internet, such that they appear to discuss their involvement more often. Some of the tweets posted by participants in long term studies showed that these participants have a stable relationship with the program. These users specifically shared their long term participant experience and feelings about the program. The word "years" appears six times in 32 tweets.

Tweets can contain search keywords but lack user membership information. 3,901 of the 4,020 (97.3%) selected tweets do not contain user participant information. Program-related accounts posted most of these tweets and tended to follow one of two patterns. The

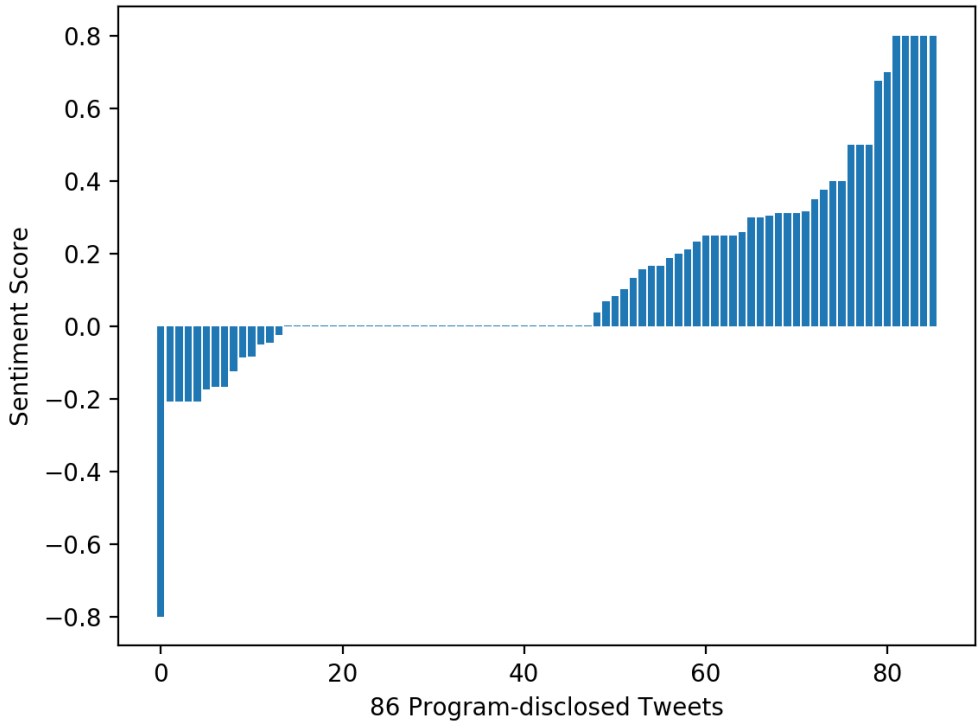

Figure 3.8: Sentiment analysis of 86 self-disclosed tweets.

first is to call for volunteers: "Come and join the xxx research program." The second is a thank you message to their participants: "xxx participants finished sequencing! Thank you, everyone, for taking part in our research!". On the other hand, tweets posted by users revealed their interest or concern about the program. For example: "I am interested in join the xxx study, but I am worried about my privacy." In general, it was observed that people are willing to join cohort studies and make their contribution, but a concern of privacy protection is an impediment. For example, 16 tweets talked about the participants' email address disclosure problem of Personal Genome Project UK.

Potential Risk of Membership Disclosure. Based on this analysis, we partitioned the risk of membership disclosure into three types: membership disclosure, identity disclosure and attribute disclosure. Here, we will discuss these privacy threats and illustrate how they relate to the specific population we studied.

Table 3.6: Year launched, number of participants and the number of tweets disclosed for the 15 cohorts.

	Study Cohort	Disclosing Tweets	Year Launched	Participants
1	Personal Genome Project	26	2005	10,000
2	100,000 Genomes Project (Genomics England)	16	2012	100,000
3	Black Women's Health Study	12	1995	59,000
4	Raine Study	11	1989	2,868
5	UK Biobank	10	2007	500,000
6	All of Us Research Program	10	2017	20,000
7	Qatar Biobank	5	2012	20,000
8	Nurses' Health Study	4	1976	280,000
9	Australian Longitudinal Study on Women's Health (ALSWH)	3	1996	57,000
10	1970 British Cohort Study (BCS70)	3	1970	17,000
11	Framingham Heart Study	2	1971	14,000
12	Millennium Cohort Study	2	1991	200,000
13	Million Women Study	2	1996	1,319,475
14	National Child Development Study	2	1958	17,415
15	Human Longevity	1	2013	N/A
	Summary	**109**		

The problems induced by membership disclosure are best illustrated with several examples. First, imagine that a volunteer has disclosed his/her membership in some research program. An attacker collects the volunteer's demographic information (e.g., residential geographical area, gender, and date of birth) from the social network (e.g., the user's Twitter profile) and links this information to the de-identified participants' records published by the research program. If a unique linkage to a record transpires, then the attacker has achieved an identity disclosure [111]. If multiple records are linked to the user, but they share the same (or similar) sensitive attribute value(s), then a successful attribute disclosure attack [144] has been perpetrated. Even if their values for the sensitive attribute are different, the attacker can guess the right one with some confidence. By contrast, previous high-profile attacks are limited in that they need to make assumptions about whether a targeted individual is indeed in a dataset. Thus, their claimed attacking powers need to be discounted by the prior probability that a targeted individual has been selected from a broader population [145]. In our scenario, the attacker is confident that the targeted individual is in the dataset. As a consequence, the discovery of membership significantly increases the likelihood of a successful attack. This attack adds significant power to all the previous attacks, which include the following:

1. Membership Disclosure. As noted earlier, the action of disclosing one's membership leaks some of the users' sensitive information. For instance, a project may be disease-specific, such that all of the participants have the same diagnosis. Similarly, some of the users join a study because they, or their relatives, have a rare disease. When they post such information online, their health information is leaked as well.

2. Identity Disclosure. By sharing membership and other personal information over social media, users can be identified. This can be accomplished by collecting self-disclosed users' personal information from their profile, such as their real name, race, gender, residence, education level, and occupation. To illustrate this issue, we randomly selected ten users and inspected their Twitter profile. It was found that nine out of ten users revealed their real face as their avatar, eight shared their location to a specific city, seven talked about their occupation or education level in their biography, six used their real name as their account name and two users made their date of birth public. With such information on hand, an attacker could find the person through a people search website, such as Intelius.com or InstantCheckMate.com. Moreover, program-disclosed individuals are more readily identifiable because the story shared by programs often contains detailed information about the storyteller. In this case, we learn the volunteer's personal information from the story, as well as their health information.

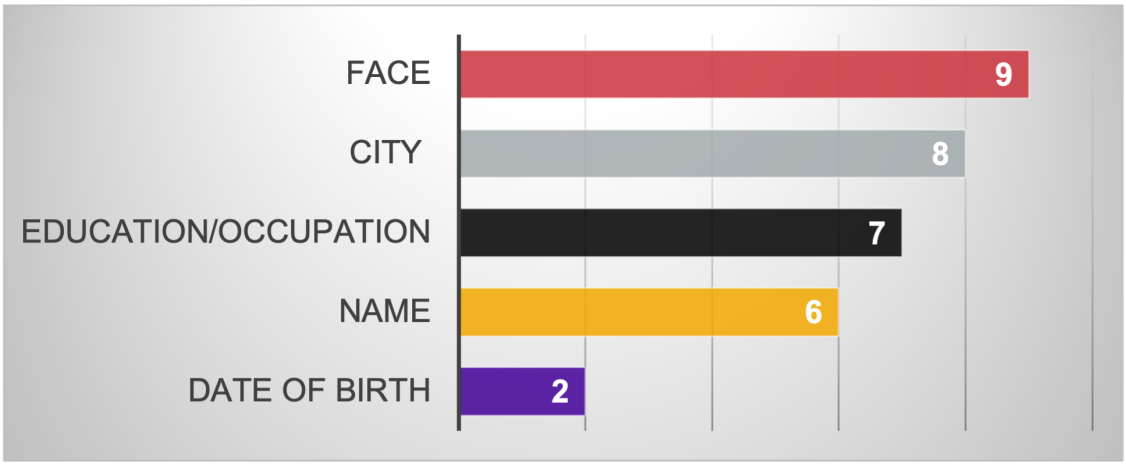

Figure 3.9: Demographics obtained from the Twitter profiles of 10 randomly selected self-disclosure accounts.

3. Attribute Disclosure. Research programs may publish their data to the public or share it with researchers in a de-identified fashion. However, if a malicious attacker has access to the cohort data, along with additional information about the self-disclosing participant (collected from the user's social media profile), then the attacker can use such information as quasi-identifiers to link to the participant's record in the cohort database. As mentioned earlier, Sweeney and colleagues [107] showed that they could identify more than 40% PGP participants using their ZIP code, gender, and date of birth, and obtain participants' sensitive information, such as medical conditions and DNA sequence.

3.5 Discussion

3.5.1 Principle Findings

This investigation illustrates that an individual's membership in a biomedical research study can be disclosed in social media in several ways. We uncovered tweets that revealed the membership over 100 participants in 15 research programs. Approximately 80% of the tweets correspond to user self-disclosure, while the remaining correspond to disclosures made by the program organizer. We found that 39 out of 86 (45.3%) self-disclosed users have a positive attitude towards joined research project. The terms "proud", "interest", and "love" were communicated by multiple self-disclosers. The personal information reported in the profiles of the social media users increased the risk of identification, which increases the likelihood that an attacker could link to their record in a de-identified dataset about the cohort, leading to further privacy intrusions, such as the re-identification of genomic information. A program may disclose participants membership when they introduce volunteer and share their story to the public as a way to increase program influence and recruit more participants. These stories may contain personal information and sensitive health information about the volunteer.

3.5.2 Limitations

Despite the findings of this study, there are limitations to this investigation we wish to highlight, as we believe they provide opportunities for improvement and extension.

First, in the membership detection attack using PheWAS data, we assumed that the prevalence of each PheWAS code is independent of other codes. However, this is unlikely to be the case in practice, such that the LR test employed by the attacker may not be as accurate as suggested. Moreover, if a targeted individual lacked an indication of a certain diagnosis, we assumed they did not have it. Yet, in reality, a lack of a diagnosis may not be definitive, such that the model may need to focus only on positively documented diagnoses and neglect those that fail to be indicated. In the standard LR test setting, the pool should be a part of the reference set (i.e., it is anticipated that the pool is a biased selection from the reference population). But in the data we collected, eMERGE-PGx is not a subset of others. However, the magnitude of eMERGE-PGx is relatively small (only 0.4% of all patients who visited the VUMC), such that it is safe to assume that the affiliation between the pool and reference will have minimal impact on the result of this investigation. In our experiment, we cycle over an entire range of all LR scores to find the best threshold. But in practice, the attacker is unaware of which threshold is the best one. This will reduce the feasibility of the attack.

Moreover, our search procedure in investigating the membership disclosure is somewhat *ad hoc*, such that we failed to detect some tweets about membership disclosure that lack certain words (e.g., participant or volunteer). We studied disclosure behavior only on Twitter, but the same problem may exist in other social platforms, such as Facebook and Instagram. A comprehensive study on additional popular social platforms is needed. The current tweet identification process requires a final manual review, but it is likely that, with enough instances of disclosure, an automated approach for discovery of such tweets could be developed. At the same time, we believe that if automated approaches can be designed to detect such disclosures, they may also be oriented to assist individuals and program

managers to recognize when disclosure is happening inadvertently. It may be that such detection and reflection of the potential risks of such actions may change decisions to reveal such information, and at least lead to more informed decision making.

3.5.3 Conclusion

Mitigating the risk of membership disclosure is not an easy problem to solve. At the end of this chapter, we wish to offer several possible strategies that may warrant consideration. First, given this threat, research programs could inform participants about the risk of membership disclosure and make it clear that if self-disclosures are made that their privacy may not be guaranteed. At the same time, research programs should inform participants of such threats when asking whether they can share information about participants (e.g., through stories). Alternatively, the program could consider sharing stories without mentioning the volunteer's real name or quasi-identifiable information.

Table 3.7: A summary of the datasets studied in the membership detection investigation.

Dataset	Population	Num. PheWAS Codes	Summary Statistics	Individual-level Records	Sample Proportion in eMERGE-PGx
eMERGE-PGx	8,173	579	Yes	Yes	-
VUMC SD	2,155,348	564	Yes	Yes	11.0%
KPW	2,446,230	452	Yes	No	12.1%
NW	1,602,402	452	Yes	No	8.9%

CHAPTER 4

Understanding Online Sharing of Genetic Testing Results on Reddit

In this chapter, we study the behavior of sharing direct-to-consumer genetic testing results in an online environment. Initially, genetic testing consumer shared their testing results anonymously, but more recently, they have included face images when discussing their results. Various studies have shown that sharing images on social media tends to elicit more replies. However, users who do this forgo their privacy. When these images truthfully represent a user, they have the potential to disclose that user's identity. We investigate the face image sharing behavior of direct-to-consumer genetic testing users in an online environment to determine if there exists an association between face image sharing and the attention received from other users. Specifically, we collected over 15,000 posts from the r/23andme subreddit, published between 2012 and 2020. Face image posting began in late 2019 and grew rapidly, with over 800 individuals revealing their faces by early 2020. The topics in posts including a face were primarily about sharing, discussing ancestry composition, or sharing family reunion photos with relatives discovered via direct-to-consumer genetic testing. On average, posts including a face image received 60% (5/8) more comments and had karma scores 2.4 times higher than other posts.

4.1 Introduction

The cost of genome sequencing has steadily decreased over time [146], which, in turn, has enabled the emergence of publicly consumable direct-to-consumer genetic testing (DTC-GT) services [147]. DTC-GT allows consumers to learn about their own genetic information without an initial consultation with a healthcare provider. The number of people who have participated in DTC-GT has increased dramatically, growing from 12 million in January 2018 to 26 million in January 2019 [148]. As of late 2021, the two largest DTC-GT companies, AncestryDNA and 23andme, had amassed over 20 million and 12 million

clients respectively [149]. Recent studies indicated that people pursue DTC-GT for various reasons, but primarily to learn about their ancestry and to discover or confirm kinship [4, 150].

As DTC-GT services have grown in popularity, consumers have increasingly relied upon online social platforms to discuss and share their test results (though not always the raw genome sequences) [16]. One particularly notable platform is Reddit, an online content rating and discussion site where users can create different subreddits based on specific topics of interest. r/23andme is one of the most popular subreddits related to DTC-GT, with more than 81,400 subscribers as of May 2022. In r/23andme, users discuss a wide range of topics related to genetic testing, including testing services, test results, explanations and interpretations, and stories about what happened after undergoing testing (e.g., health-related decisions) [16].

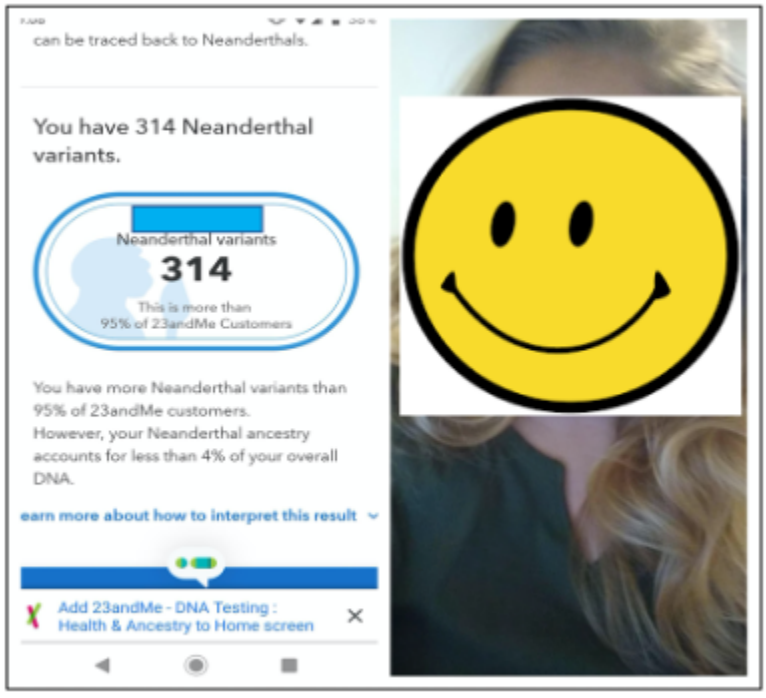

Figure 4.1: An example of a face image posted on the r/23andme subreddit: report together with a face image and testing results.

When r/23andme users share their results for discussion, instead of simply typing text, some users attach a screenshot of their DTC-GT result page (e.g., the ancestry com-

position). Since Reddit is a virtual online community where users generally rely upon pseudonyms for communication, such screenshots of results typically do not contain a user's real name. Therefore, even when users share and discuss their DNA test results, this subreddit has historically been a community with a culture of anonymity.

However, in 2019, r/23andme users began attaching personal images to their posts. Figure 4.1 presents such an example with a screenshot of the user's DTC-GT result page on the left and the full-face image of this user on the right. The actual face and name are obscured; however, the data exist in the public domain. This image sharing movement towards revealing one's face directly affects personal privacy [151, 152]. Although these posts used pseudonyms, face image posting in online environments constitutes a knowing decision to give up one's privacy. For example, other users may utilize these face images to determine a user's identity, relying, in part, on the rapid development and deployment of modern face recognition [153] and identity detection systems [154]. This is a concern because identity disclosure may lead to various negative consequences for individuals, including identity theft [155], discrimination [156], and threats to one's safety [157]. Since Reddit is a public platform, a user's posts and face images are readily accessible, making an identity disclosure attack feasible with little cost [158].

Though users may be aware that revealing one's face likely compromises their privacy, it is unclear why they choose to do so. Various investigations into behavioral psychology and economics show that some people waive their privacy rights in exchange for a service that they value [159]. Thus, we hypothesize that r/23andme users may receive more attention by publishing more personal or revealing information. This is supported by the findings in other social platforms as well. For instance, Tweets with photos can boost retweets by 35% [160], Instagram photos with faces are 38% more likely to receive likes and 32% more likely to receive comments [161]. However, unlike Twitter or Instagram, the DTC-GT forum provides an anonymous environment for users to share and discuss sensitive personal genetic information. Therefore, it is worth investigating whether the same privacy-service

exchange hypothesis holds in this forum. To formally test our hypothesis, we investigate the following questions:

RQ1: What are the topics communicated in the natural language of posts with face images?

RQ2: Is face image posting associated with the attention that a post receives?

To address these questions, we first collected posts from the r/23andme subreddit and categorized them into three types: post with text only, post with face images, and post with images not containing a face. Then, we measured the temporal posting trends regarding the type of post. Next, we applied topic modeling to compare the difference of primary topics associated with three types of posts. Finally, we performed a regression analysis to infer the association between the attention that a post receives, in terms of votes, comments, and the type of the post, and whether the post contained a face image.

4.2 Related Work

Natural language processing techniques have been applied to various healthcare applications [162]. Considering healthcare-related social media studies as an example, Liu and colleagues [20] analyzed the association between weight loss progress and Reddit users' online interactions; Klein and colleagues [21] relied upon Twitter data to identify potential cases of COVID-19 in the United States; and Ni and colleagues [37] compared the attitudes of users of four different social platforms towards #GeneEditedBabies. For DTC-GT, most of the investigations focus on consumer motivation [163], health implications [164], and ethical implications [165], but only a handful of them have considered the disclosure of test reports over social platforms [151, 166]. Most previous studies using social media data focused solely on mining knowledge from the text. In this chapter, by taking image posting into consideration, we assess the behavior of personal image sharing on this DTC-GT forum.

In this chapter, we analyzes the association between face image sharing and attention

achieved in an online setting, the latter of which may incentivize users to sacrifice their privacy in exchange for the benefit of a social response. This observation, however, does not imply that attention is undesirable in all cases, as several studies have shown that social engagement is beneficial to an individual's physical and mental health. For instance, in a large online breast cancer forum, Yin and colleagues [167] found that the volume of online interchange is positively associated with patient treatment adherence. Pan and colleagues [168] found that receiving replies can benefit online participants in depression forums. Naslund and colleagues [169] analyzed the benefits and risks of using social media as a potentially viable intervention platform for offering support to persons with mental disorders. Thus, the perceived benefits an individual receives from a service typically outweighs the perceived privacy risks in the near term. but given that privacy concerns tend to be realized at a later point in time [170], Reddit may wish to consider warning users about the potential negative consequences of their actions.

4.3 Data and Methods

Figure 4.2 provides an overview of the research pipeline, which consists of two primary steps. The first step involves data collection and categorization, where we collected the posts on the r/23andme subreddit and extracted those with a face image using face recognition software. The second step focuses on analysis. Specifically, we first conducted an exploratory analysis to investigate the temporal posting trends and then leveraged topic modeling to infer the themes communicated in these posts. Finally, we performed a regression analysis to determine whether including a face image in a post was associated with the attention it received. In this chapter, we characterized attention by the number of comments and the karma score that a post received from other online users. The karma score in Reddit is defined as the number of upvotes minus the number of downvotes, indicating the popularity of a post.

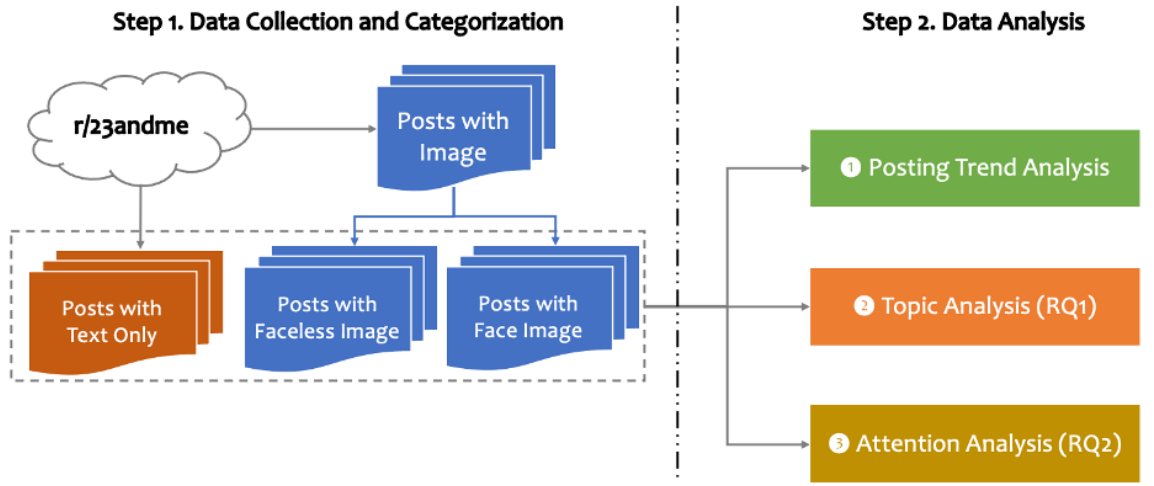

Figure 4.2: An overview of the research workflow for posts in the r/23andme subreddit.

4.3.1 Data Collection and Categorization

To collect data from the r/23andme subreddit, we first gathered the IDs of all posts (i.e., submissions) and comments using pushshift.io. We then applied the Python Reddit Application Programming Interface (API) Wrapper package (version 6.3.1) to extract data from the Reddit for each ID. Specifically, we collected all posts and comments published on r/23andme between December 31, 2012, and January 31, 2020. Each collected post contains the following information: 1) author identifier, 2) post title, 3) post text body, 4) image URL (if there is an image in the post), 5) comments on the post, 6) publication date, and 7) karma scores of the post and affiliated comments.

We downloaded the images for the posts containing an image URL and applied the face-recognition Python package (version 1.3.0) [171] to classify images into 1) images with a face and 2) images without a face (i.e., faceless images). To assess the accuracy of the face detection algorithm, we randomly selected 100 images from each group and manually examined the quality of classification. We found that seven faceless images were classified as face images, indicating a false positive rate of 7% (7/100), while two face images were classified as faceless images, indicating a false negative rate of 2% (2/100). To achieve 100% precision, we manually reviewed all the images in the face group and

relabeled the misclassified images. Due to a high true positive rate of 98% (98/100) and the large volume of the faceless images (3,865), we did not perform a manual review step for the set of faceless images. As such, we categorized all of the collected posts into three types: 1) text-only posts; 2) posts with faceless images; and 3) posts with faces (e.g., Figure 4.1), which resulted in three types of users.

4.3.2 Data Analysis

To describe face image posting behavior, we compared the face posts with the other two types of posts along three perspectives: 1) posting temporal trend, 2) post theme, and 3) the attention that a post received from other users, in terms of the number of comments and karma score.

Topic Analysis. To examine the thematic differences between the three post types, we applied topic modeling [172] to the post title instead of to the post body. This is because 41.1% (6404/15596) of the post had an emtpy text body. We first tokenized the data and removed all punctuation. Next, we lemmatized words into their base forms (e.g., "walks" becomes "walk") using the nltk Python package (version 3.3). We also replaced personal pronouns, such as "we", "she", and "they", with the symbol "-PRON-", and the numbers with the word "datum". We then applied Latent Dirichlet Allocation (LDA) [28], as implemented in the gensim Python package (version 3.8.1), to extract topics. Since LDA is an unsupervised learning model, we calibrated the number of topics for the optimal model based on coherence score, which measures the pairwise word semantic similarity in a topic. To do so, we ran LDA models with 2 to 20 topics (using a step size of 2) on the set of lemmatized words and selected the topic number that achieved the largest coherence score. Finally, to demonstrate the quality of topic modeling, we used t-distributed stochastic neighbor embedding (t-SNE) [173] to cluster topics and display the results in a two-dimensional representation.

Regression Analysis. We investigated two types of associations. First, we considered

the association between a faceless image post and the attention it received. Second, we considered the association between a face post and the attention it received. We used the number of comments and the karma score to measure the attention of a post. Since these numbers are non-negative count variables, we applied a negative binomial regression to infer the association [174].

Given that posts published earlier may be read by more readers and, thus, induce more comments and votes, we included the number of days a post had been published as a control variable. In addition, posts on different topics might receive different levels of attention. To reduce the effects of post topic, we incorporated the topic distribution of each post as an additional set of control variables. During model fitting, we dropped one topic (T4, see below) to address collinearity.

Moreover, user activity might affect the popularity of their posts. For example, posts from active users may receive more attention. To reduce the impact of user activity, we incorporated the number of posts and the number of comments of each user as an additional set of control variables. We utilized the implementation of negative binomial regression in statsmodels Python package (version 0.11.1) to fit models for the karma score and the number of comments, separately. We reported the features that achieved statistical significance at the .001 level.

4.4 Results

We collected 15,596 posts and 188,843 comments, which were published by 20,883 users, between December 31, 2012 and January 31, 2020. Among the collected posts, 24.8% (3818 / 15596) posts contained faceless images, while 5.4% (849 / 15596) of the posts contained face images.

4.4.1 Exploratory Analysis

Temporal Trend. In Figure 4.3, the graph to the left depicts the temporal post trend on a monthly basis. It can be seen that the r/23andme subreddit exhibited relatively low

activity until 2017, after which time the number of monthly posts grew rapidly. Image posts (with and without a face) became popular after 2018. The graph to the right of Figure 4.3 shows the quarterly growth rate of the number of posts. The green dotted line indicates that, since 2019, the number of face posts exhibited a rapid increase, with a growth rate that surpassed the growth rate of all posts (represented by the blue line) and image posts (represented by the orange dashed line). Notably, we find that posting rates for all three type of posts increased rapidly after 23andme's major promotions (Amazon Prime Day and Black Friday), which is consistent with the findings of Yin et al [16].

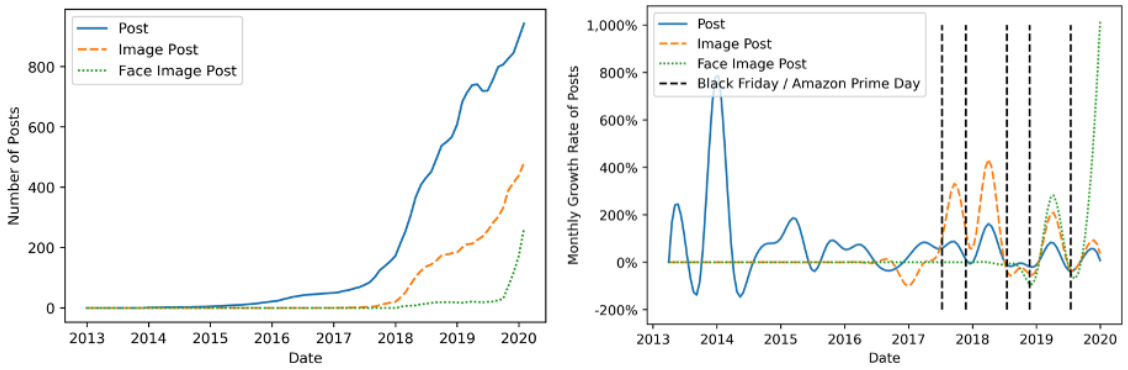

Figure 4.3: Smoothed temporal trends of three type of posts: number of posts published per month (left); and quarterly growth rate of number of posts (right).

Attention to Posts. In Figure 4.4, the left graph depicts boxplots for the number of comments per post for each post type. Face posts received the most comments, followed by posts not containing a face. The median number of comments for text-only posts was 6, but the median increased to 9 for posts with faceless images and 13 for posts with face images. The right graph in Figure 4.4 shows the karma score by post types. The face posts received the highest median karma score of 34, followed by faceless posts, which had a median karma score of 13. By contrast, the median karma score for text posts was only 4. One-way ANOVA tests for comments and karma scores indicated that the differences are statistically significant ($P < .001$). For presentation purposes, we removed posts with more than 80 comments or karma scores greater than 150 (3% of the data) in Figure 4.4.

User Activity. We measured the user activity in terms of the number of posts and

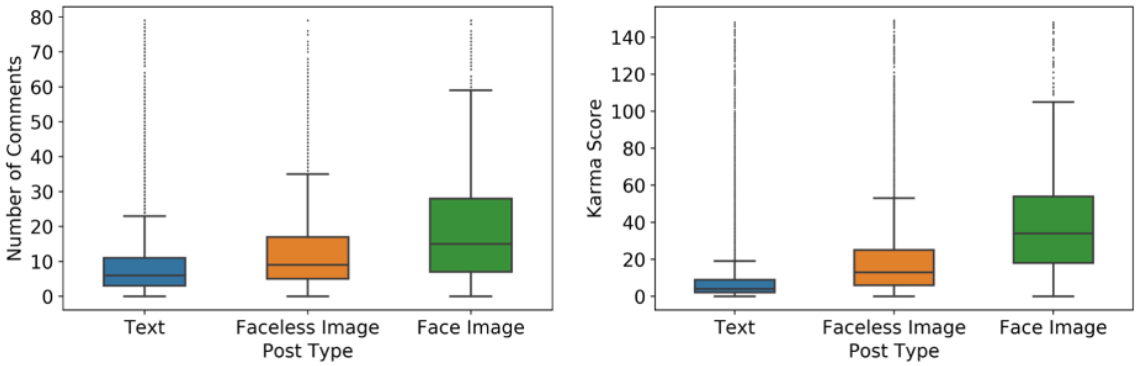

Figure 4.4: Attention of three types of posts: number of comments per post (left); and Karma score per post (right).

comments. 26.8% (2442 / 9114) of the users posted faceless images, while 8.5% (774 / 9114) of the users posted face images. In Figure 4.5, the left graph shows that the median number of posts for all three user types was 1. However, the third quartile of users who posted images (with or without a face) was 2. This suggested that, on average, authors who posted images (with or without a face) had more posts than authors who posted text only. The right graph in Figure 4.5 depicts the number of comments posted for each author type. The users who posted face images wrote the most comments, with a median of 8. The median dropped to 6 for users who posted images not containing faces. For users who posted text only, the median number of comments was substantially lower at 3. The results of one-way ANOVA tests for the number of posts and the number of comments showed that the differences are statistically significant ($P < .001$). For presentation purpose, we removed users published more than 10 posts or 50 comments from figures, accounting for 4.4% of the total number of users in Figure 4.5.

4.4.2 Topic Analysis

To decide the number of topics within the post corpus, we ran LDA models with 2 to 20 topics using a step size of 2. Figure 4.6 illustrates the change in coherence score as a function of the number of topics. We selected 10 as the number of topics, as it achieved the highest coherence score (0.391).

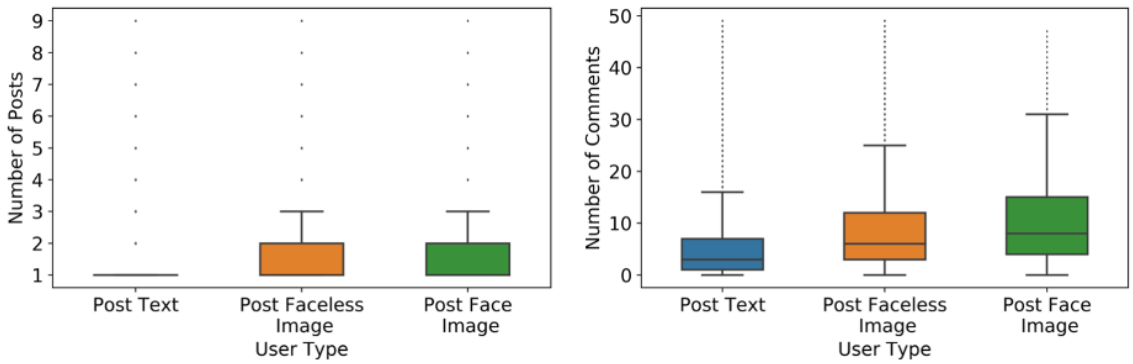

Figure 4.5: Number of posts per user (left) ; and number of comments per user (right) for three types of users: users who post text only, users who post a faceless image, and users who post a face image.

Table 4.1 shows the 10 inferred topics, their most relevant words and the topic distribution. The most relevant words were ranked based on their marginal distribution within a topic and displayed in descending order. The topic distribution was calculated as the percentage of posts belong to the topic. Based on the relevant words and posts with the highest probability for each topic, we further grouped the 10 topics into three categories: 1) Ancestry Composition; 2) Kinship and Family Discovery; and 3) General Questions about Genetic Testing.

Figure 4.8 displays the distribution of the three post categories and 10 topics in a 2D t-SNE scatterplot. Instead of showing all the posts in one figure, we selected 10% (1,587/15,596) of the most relevant posts to their dominant topics (the topic with the largest proportion). All the selected posts achieved a relevance score (the relevance between the post and its dominant topic) higher than 0.155. In general, the 10 clusters in Figure 4.8 are well separated, but cluster T5 and cluster T6 were close to each other, which suggests these two topics have a high degree of similarity.

Ancestry Composition included four topics: T1, T2, T3, and T4. Posts in this category focused on the presentation and discussion of the ancestry composition testing results. The four topics captured the ancestry information, which communicate a user's race, continental origin, and nationality. The following posts are examples of this category:

Category	Topic ID	20 Most Relevant Terms	Topic Distribution
Ancestry Composition	T1	European, -PRON-, result, Italian, Irish, British, surprise, Jewish, white, Chinese, broadly, bit, eastern, Ashkenazi, surprised, Scandinavian, give, eye, lot, surprising	11.6%
	T2	-PRON-, ancestry, German, guess, French, make, post, heritage, year, ethnicity, grandmother, common, grandparent, explain, mega-thread, feel, polish, Canadian, confused, wrong	7.9%
	T3	result, -PRON-, expect, finally, back, ancestor, interesting, pretty, AncestryDNA, bear, confidence, recent, location, Filipino, cool, guy, live, thought, Finnish, big	9.1%
	T4	American, Asian, African, native, Mexican, people, south, percentage, region, Neanderthal, gene, high, part, Spanish, unassigned, east, north, variant, trace, add	10.6%
Kinship and Family Discovery	T5	-PRON-, family, today, close, tree, understand, worth, info, don, trait, history, link, happen, picture, excited, love, list, connection, inherit, risk	6.5%
	T6	–PRON-, find, dad, half, mom, father, cousin, mother, side, sister, adopt, brother, great, sibling, grandfather, full, grandma, biological, aunt, figure	9.2%
General Questions	T7	kit, long, time, extraction, wait, timeline, genetic, day, receive, sample, analysis, week, testing, step, send, batch, fail, information, work, stick	14.2%
	T8	-andme, ancestry, datum, health, raw, accurate, GEDmatch, MyHeritage, good, DNA, upload, compare, site, comparison, land, data, service, difference, WeGene, interpret	11.0%
	T9	DNA, test, relative, question, parent, report, share, -PRON-, phase, show, generation, relate, computation, person, unexpected, noise, mystery, relationship, account, number	9.7%
	T10	result, update, beta, haplogroup, match, maternal, change, paternal, chromosome, map, mixed, chip, Puerto Rican, Korean, lose, comment, late, original, Romanian	10.2%

Table 4.1: The topics inferred from r/23andme subreddit.

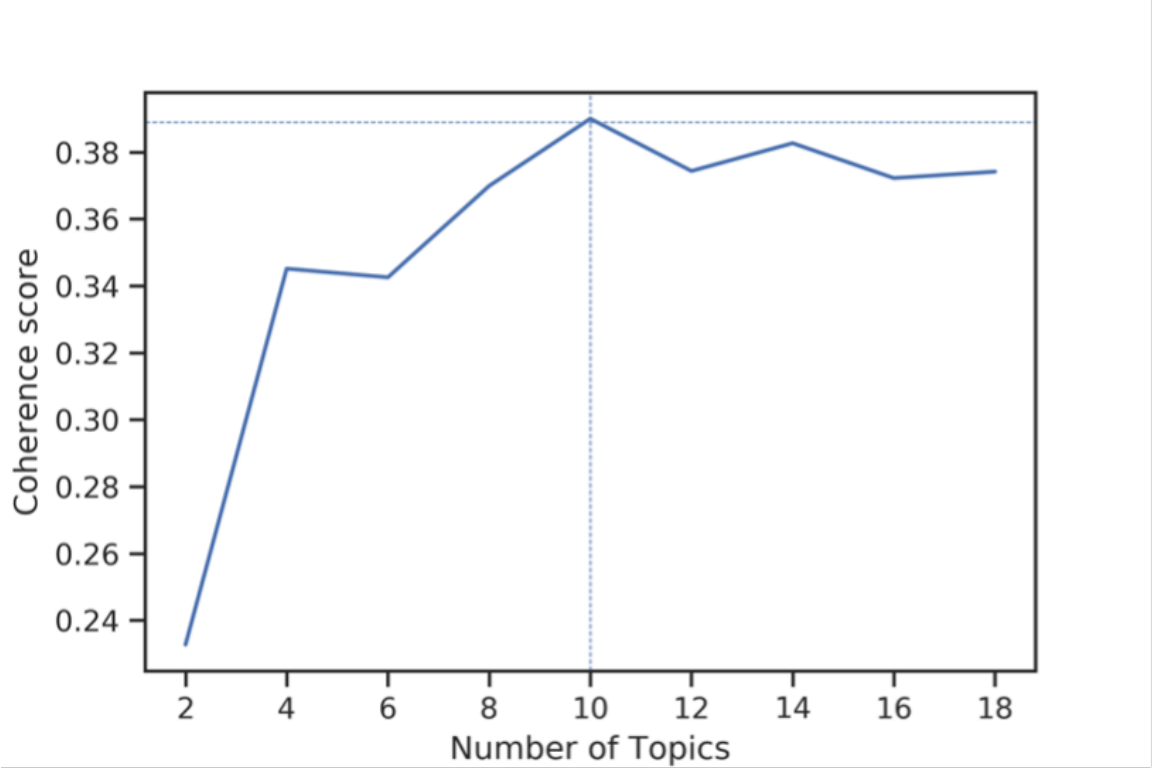

Figure 4.6: Coherence score as a function of the number of topics.

1. "So I'm a lot less British than I thought, and a lot more Swiss." (T1)
2. "Any guesses on my friend's ethnicity? He thinks he's French/German, English, and maybe some Slavic." (T2)
3. "Born and raised in Manila, grew up thinking I was 100% Filipino. A bit shocked at my results." (T3)
4. "Found out I am East Asian and Native American but I have northern Asian and Native American so high." (T4)

Kinship Finding and Family Discovery was communicated in T5 and T6. Specifically, T5 communicated the discovery of ancestors and distinct relatives, where it can be seen that "family" and "history" were often used. In T6, words such as "find", "dad", and "sib-

lings" showed that this topic focused on the findings of immediate family members. The following examples highlight this observation:

1. "My cousin did the DNA test and connected us to our great grandmother's family!" (T5)
2. "Found out I have about a dozen cousins I didn't know about." (T6)
3. "On my account apparently my mom and her twin sister are both my moms." (T6)

General questions related to DTC-GT were communicated in Topics T7, T8, T9, and T10 discussed related to DTC-GT. Specifically, the posts in T2 mainly asked about testing service progress. Words such as "time" and "wait" were highly weighted in this topic. The posts in T7 were mainly about the comparison between DTC-GT companies. There were mentions of companies, such as "MyHeritage", "23andme" and "WeGene". Topic T8 covered posts about understanding, or questions about, the test result report. The posts in T10 mainly discussed the upgrade of the genetic testing algorithm and the subsequent changes in testing results. Words such as "beta", "update" and "change" were highly weighted. Relevant examples include:

1. "Is my kit moving slow? It took 2 weeks to be marked as "arrived" after tracking showed it was delivered." (T7)
2. "23andMe vs WEGENE - uploaded 23andMe raw data to WEGENE and here are the differences." (T8)
3. "What is a likely relationship if the shared DNA is 1610 centimorgans across 80 segments?" (T9)
4. "Beta update v5.2 should now be available to all earlier chip (pre-V5) users, when opting into the Beta program" (T10)

Figure 4.7 presents the topic distribution for each type of posts. In this figure, topics are

arranged according to categories, where $*$ indicates that the pairwise differences between the three post types for the topic are statistically significant according to a one-way ANOVA with post-hoc Tukey HSD tests, $P < .001$. The one-way ANOVA tests showed that there are statistically significant differences between the means of the three post types for all 10 topics ($P < .001$). From the figure, it can be seen that face posts are more likely to communicate Ancestry Composition (topics T1, T2, T3, and T4) and Kinship and Family Discovery (topics T5 and T6), while text posts are more likely to ask General Questions (topics T7, T8, and T9). For T10, a topic about 23andMe algorithm upgrade, it can be seen that faceless image posts are more likely to communicate this topic, followed by text posts and then face image posts. This may be because users tended to post screenshots of the results before and after the algorithm upgrade for easy comparison.

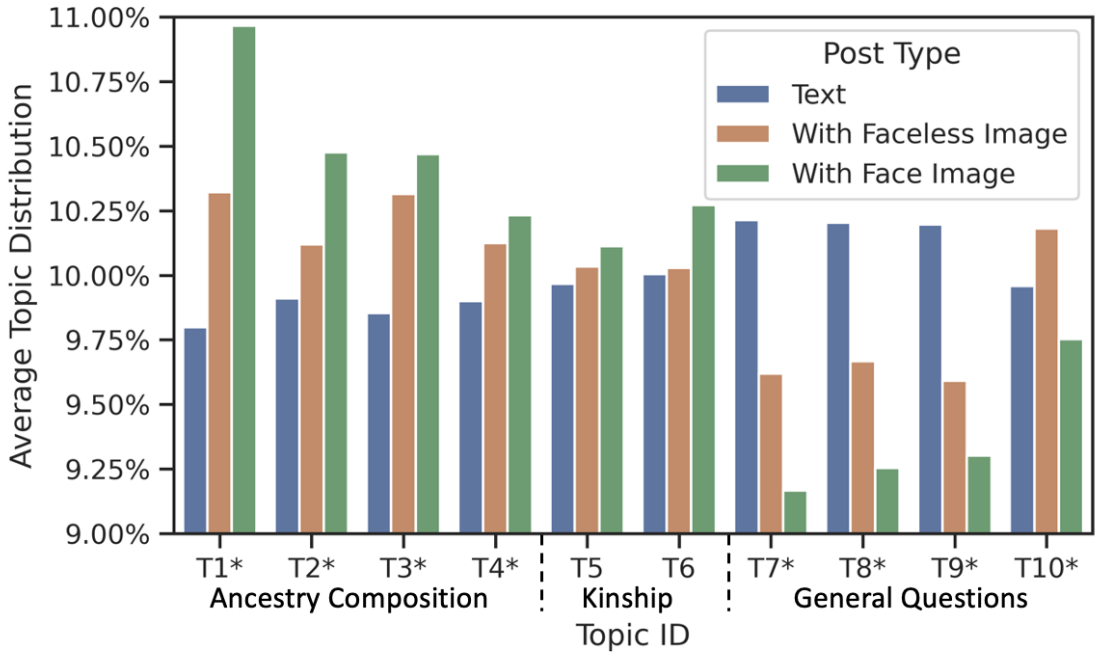

Figure 4.7: The prevalence of topics in each post type.

4.4.3 Regression Analysis

Table 4.2 summarizes the results of the negative binomial regressions. The two regressions $R_{image \rightarrow comment}$ and $R_{image \rightarrow score}$ indicate the association between the number of

comments, karma score, and whether the post contained (faceless or face) images. Image posting exhibited statistically significant positive associations with both dependent variables, which suggests that image posts received more attention than text-only posts.

Table 4.2: Results of the regression analysis relating post type to comments and karma score.

Negative Binomial Regression	**Dependent Variable**	**Independent Variable**	β	**Z**	**std**	**P-value**
$R_{image \to comment}$	Number of Comments	Posting Image	.152	6.41	.024	1.43×10^{-10}
$R_{image \to score}$	Karma Score	Posting Image	.618	12.35	.050	4.70×10^{-35}
$R_{face \to comment}$	Number of Comments	Posting Face Image	.451	10.21	.044	1.85×10^{-24}
$R_{face \to score}$	Karma Score	Posting Face Image	.760	9.64	.079	5.65×10^{-22}

In $R_{face \to comment}$ and $R_{face \to score}$ tests, we selected 4,717 image posts and assessed the association between the number of comments, karma score and whether the image contained a face. Face image posting exhibited statistically significant positive associations with both dependent variables, which indicates that face posts received more attention than faceless posts. When comparing the $R_{image \to comment}$ and $R_{face \to comment}$ tests, it was observed that posting a face image achieved a more positive impact on receiving comments. A similar result was obtaned when comparing the $R_{image \to score}$ and $R_{face \to score}$ tests.

In addition, there were two notable findings with respect to the control variables. First, the (log transformed) number of published days exhibited a negative association in the $R_{image \to comment}$ and $R_{image \to score}$ tests ($\beta_{image \to comment} = -.09$, $\beta_{image \to score} = -.26, P < .001$). Second, topic T8 (comparison of DTC-GT companies comparison) had a negative association in all four tests ($P_{image \to comment, face \to comment} < .001$, $P_{image \to score} = .003$,

$P_{face \to score} = .013$), while topic T7 (testing service progress) showed a negative association in the $R_{image \to score}$, $R_{face \to score}$, and $R_{face \to comment}$ tests ($P_{image \to score} < .001$, $P_{face \to score} = .003$,$P_{face \to comment} = .041$). The negative association between T7, T8 and face posting further justifies our previous finding: the topics in posts including a face were less likely to correspond to a general question about DTC-GT. We report the coefficient, z-score and p-value for for $R_{image \to comment}$ and $R_{image \to score}$ in Table 4.3, and report the same set of statistics for tests $R_{face \to comment}$ and $R_{face \to score}$ in Table 4.4.

Table 4.3: Summary of negative binomial regression results between posting image and the number of comments and karma scores.

Dependent Variable	Num. Comments			Karma Score		
	coef.	***Z***-score	***P***-value	coef.	***Z***-score	***P***-value
Independent Variable						
Image / No Image	0.152	6.412	**<.001**	0.619	12.353	**<.001**
Control Variables						
T1. European Ancestry	0.039	0.902	.966	3.364	1.763	.078
T2. German Ancestry	3.053	3.124	.002	6.497	3.139	.002
T3. Ancestor	-1.871	-1.900	.057	2.469	1.187	.235
T5. Family Tree	-1.463	-1.466	.143	11.137	5.286	**<.001**
T6. Find Dad / Sibling	1.848	2.300	.021	17.460	10.271	**<.001**
T7. Kit Wait Time	-1.192	-1.549	.121	-10.986	-6.741	**<.001**
T8. DTC-GT Companies	-8.986	-10.799	**<.001**	-5.280	-3.023	.003
T9. Questions	-5.632	-6.267	**<.001**	-6.781	-3.580	**<.001**
T10. Result Update	-8.307	-9.039	**<.001**	-5.946	-3.072	.002
Days Published ($\log_{10}$)	-0.090	-10.147	**<.001**	-0.263	-14.121	**<.001**

4.5 Discussion

4.5.1 Principle Findings

There are several notable findings from this investigation. First, consistent with previous studies in other social platforms [161, 175], we observed that posts with faces in

Table 4.4: Summary of negative binomial regression results between posting face image and the number of comments and karma scores.

Dependent Variable	Num. Comments			Karma Score		
	coef.	***Z***-score	***P***-value	coef.	***Z***-score	***P***-value
Independent Variable						
Face / No Face	0.450	10.207	**<.001**	0.7569	9.636	**<.001**
Control Variables						
T1. European Ancestry	-2.469	-2.335	.02	-2.984	-1.592	.111
T2. German Ancestry	-2.353	-1.983	.047	-1.698	-0.806	.420
T3. Ancestor	-2.7928	-2.331	.020	1.347	0.635	.525
T5. Family Tree	-0.318	-0.239	.811	7.6907	3.260	.001
T6. Find Dad / Sibling	0.214	0.200	.839	10.429	5.563	**<.001**
T7. Kit Wait Time	-2.344	-2.044	.041	-6.094	-2.997	.003
T8. DTC-GT Companies	-5.717	-4.563	.001	-5.4673	-2.474	.013
T9. Questions	0.090	0.061	.951	6.453	2.452	.014
T10. Result Update	-6.000	-5.018	**<.001**	-7.133	-3.374	.001
Days Published ($\log_{10}$)	-0.0364	-2.589	.010	-0.037	-1.495	.135

the r/23andme subreddit receive more attention than other posts. It is possible that the increase in attention drives the disclosure of personal information in such online environments. However, it should be noted that our investigation is not causal, and this is only a conjecture at this time. Regardless of the motivation for face image posting, it is evident that this behavior is growing rapidly within this subreddit.

Second, the 10 inferred topics from the titles of r/23andme posts appeared to fall into three categories. The posts in the first category, which covered four out of 10 topics, focus on discussing users' ancestry composition. Notably, the topics in this category were associated with a higher rate of image and face image posting. It was further observed that users invoke their face images as proof (or a counterexample) of the genetic testing results. Posts about kinship and family member discovery exhibit a moderate rate of face image sharing. When inspecting posts in this category, posts such as "finally find my half-sister",

with a group photo attached as a reunion, were more prevalent than in other categories. Finally, posts asking general questions about genetic testing, which focused on comparisons between DTC-GT companies, the progress of testing result delivery, and the upgrade of testing algorithm, exhibited the lowest rate of image sharing.

Third, counter to our expectation, it was found that the number of days a post was published was negatively associated with a post's attention. One possible explanation for this result is that Reddit archives posts older than 6 months and no longer allows commenting to archived posts. Thus, the number of comments and votes were limited for earlier posted posts. We further noticed that the topics about general questions were negatively correlated with a post's attention.

4.5.2 Limitations

Despite these findings, there are certain limitations to this work, which we believe serve as opportunities for future research. First, the face recognition package had an estimated 2% false negative rate, which means there may be about 76 (2%×3865) face images labeled as faceless images. These misclassified images might influence the accuracy of our result, although not the direction. Second, most topics inferred from topic modeling appear to be interpretable and intuitive; however, the message conveyed in topic T10 is difficult to interpret. As shown in Table 4.1, the sample words of T10 convey different kinds of information, "Puerto Rican" and "Korean" are related to ancestry composition, whereas "late" and "lost" are evidence of asking about delivery progress. In this respect, newer topic modeling techniques [176–178] or language model-based topic modeling (e.g., top2vec and BERTopic) may provide better intuition into the semantics of the posts on social platforms. Importantly, however, the quality of individual topics had little effect on our main conclusion. Since the regression analysis (using the topic distribution as control variables, Table 4.2.) and ANOVA test (without topic distribution, Figure 4.7.) yielded the same finding – a statistically significant association between face image sharing on r/23andme and user

engagement.

4.5.3 Conclusion

DTC-GT users are increasingly posting full face images with their DTC-GT results on social platforms. In this chapter, we investigated the trend in this behavior in the r/23andme subreddit to obtain insight into its potential motivation. Our findings show that such behavior began in September 2019 and experienced rapid growth, with over 849 face-revealing posts by early 2020. Furthermore, our investigation suggests that posts including a face received, on average, 60% (5/8) more comments and 2.4 times higher karma scores than other posts. Posts that included face-images were primarily about sharing and discussing ancestry composition and sharing family reunion photos with relatives discovered via DTC-GT. These findings verify our hypothesis that posting a personal image is associated with receiving more online attention, which is consistent with previous findings that people appear to be willing to give up their privacy (i.e., personal image) in exchange for a benefit (i.e., attention from others). Based on this analysis, platform organizers and/or moderators could inform users about the risk of posting face images in a direct, explicit manner and make it clear that users' privacy may be compromised if personal images are disclosed.

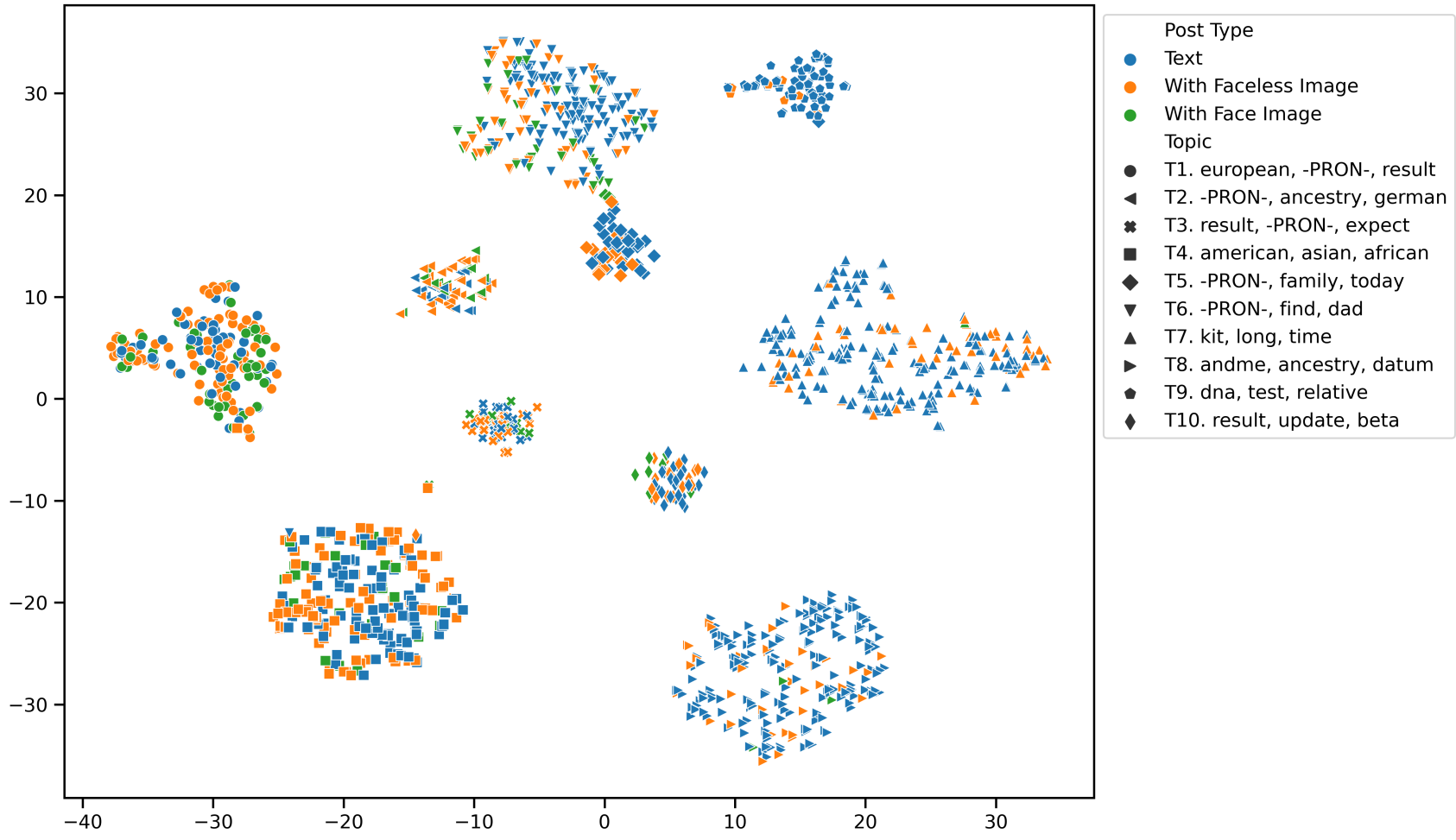

Figure 4.8: t-SNE clustering result of 1,587 selected posts in 10 topics, the markers represent posts. The relevance between selected posts and their dominant topic is greater than 0.155.

CHAPTER 5

Examining Rural and Urban Sentiment Difference in COVID-19 Related Topics on Twitter

In this chapter, we focus on a topic that affects the general population: public sentiment about COVID-19 and related topics. In this investigation, we combined word embedding models with clustering strategies to identify topics closely related to COVID-19, and relied upon the similarity between topic hashtags and opinion adjectives to infer the sentiment with respect to the identified topics. We discovered a significant difference between US urban and rural users in their sentiment about COVID-19 prevention strategies, misinformation, politicians, and the economy.

5.1 Introduction

The COVID-19 pandemic has persisted for over two years. By August 2022, more than 93 million people in the United States (US) were infected with COVID-19 with notable disparites [179]. In particular, the death rate in rural areas (325/100,000) has been significantly higher than in urban areas (248/100,000) [179, 180], a disparity that highlights the need to improve practices in prevention and control [181]. However, the path to improving the situation in rural environments is not clear, as urban and rural residents have different sentiments and attitudes towards many COVID-19 related topics. For example, it has been shown that rural residents are less concerned about the coronavirus [182], and are less willing to engage in COVID-19 related prevention behaviors [8, 183]. Moreover, political polarization influences the public's attitude and reaction to the COVID-19 pandemic [90, 184]. In this respect, it is evident that a more comprehensive investigation into the differences of opinion and sentiment between urban and rural residents about COVID-19 and related topics is needed.

To date, there have been several studies into the differences between urban and rural

sentiment about COVID-19 [8, 182, 184, 185]. However, these studies have mainly relied upon formal surveys, which are limited in their ability to shed light on the matter because they are time consuming and the findings (as well as the policies made upon them) can become stale in the face of the rapid evolvement of the situation [15]. Social media platforms have enabled people to report on their experiences and express their perspectives on COVID-19 on a wide scale. The data generated through social media have been relied upon to study various aspects of health and wellness [19, 23, 24, 37, 186], such that it is natural to hypothesize that this large and diverse collection of user-generated data provides opportunities to investigate the differences between urban and rural sentiments. In this chapter, we focus on the study of public sentiment on COVID-19 related topics using data from Twitter, one of the largest social platforms in the US, with over 200 million daily active users [17].

5.2 Related Work

While topic extraction and sentiment analysis are typical natural language processing tasks, prior research on learning sentiments about COVID-19 from social media has been limited in several ways. First, prior studies [70, 187] have relied on topic modeling techniques, such as LDA [28], to identify relevant topics from the collected social media data. However, such methods rely on document-level word co-occurrences to infer a topic distribution [33], which leads to poor topic extraction performance for noisy short text data [188]. Second, most studies applied either predefined rules [70, 84, 85, 90, 91], such as VADER [71] or machine learning models to infer sentiment from tweets. While the rule-based approaches fails to leverage the contextual information in a specific corpus, which varies by corpus, machine learning approaches [69, 79, 83, 88, 92] are hindered by their need for a non-trivial amount of label annotation and training.

Even if the labeling process can be expedited, to initiate a study with social media data, it is necessary to collect online posts with the topic of interest. The majority of previous

studies in this area applied keyword filtering to collect COVID-19 related tweets [81, 189, 190]. However, keyword filtering is hindered by an incompleteness problem that can lead to biased investigations. For example, in one vaccination opinion study [191], all of the keywords contained the prefix "vaccin-", which neglected all tweets that used the word "vax". At the same time, societal response to the pandemic is constantly evolving, with new keywords being generated at different stages. It is unlikely that one would be aware of all appropriate keywords at any point in time. For instance, in the COVID-19-TweetIDs dataset [192], the word "vaccine" was not added to keyword list until November 2021 - one year after vaccines received US Food and Drug Administration (FDA) emergency use authorization.

In this chapter, we introduce a novel approach for COVID-19 sentiment analysis. This approach begins by collecting tweets without any pre-defined keywords. To identify topics from the brief amount of text in a tweet, the approach utilizes word embedding models and clustering approach to extract topics related to COVID-19. The new approach combines lexicons and semantic information to quantify public sentiment with respect to a specific population of interest regarding COVID-19 and related topics, such as prevention, vaccination and politics.

5.3 Data and Methods

Figure 5.1 depicts the data processing and research pipeline for this investigation. It consists of three primary steps: 1) tweet collection, 2) model training and 3) sentiment analysis. The collection step involves the gathering of tweets and a designation of their urban-rural status. The model training step involves the training of multiple word2vec models based on geospatial and timing information. The sentiment analysis step consists of COVID-19 topic clustering and multi-dimensional sentiment analysis with opinion adjectives.

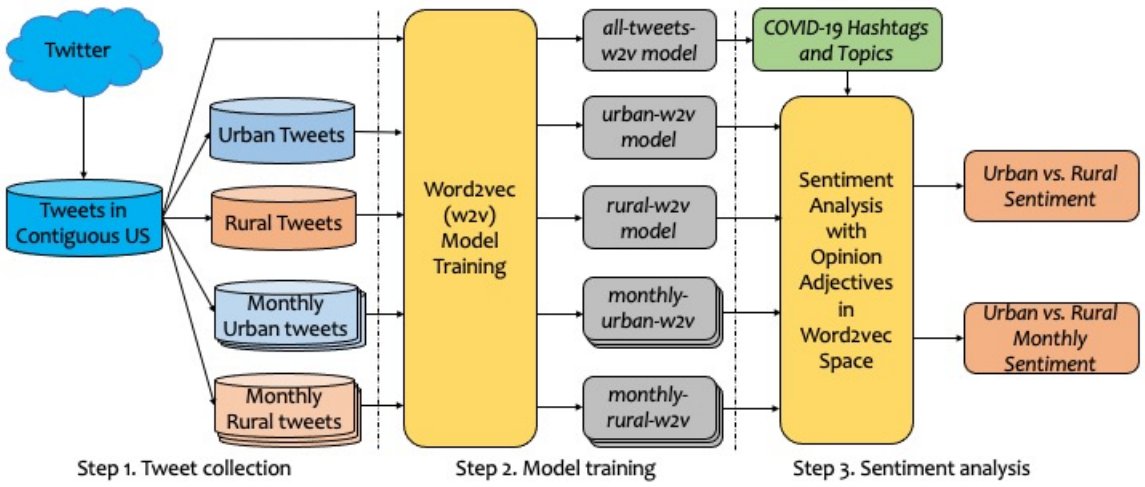

Figure 5.1: An illustration of the research pipeline.

5.3.1 Data Collection

We used Tweepy (version 3.8) to collect 407 million geo-tagged tweets posted in the contiguous US through the Twitter API streaming function between May 2020 and January 2022. A geo-tagged tweet contains location information in the form of either 1) a specific latitude and longitude or 2) a Twitter place text field. For tweets with the place field, we applied geocoding with the geopy python package (version 2.2) to obtain the latitude and longitude, which were then translated into 5-digit ZIP codes. We did not apply keyword filtering during collection, such that it is expected that the tweets are an unbiased sample of all publicly accessible US geo-tagged tweets.

Urban and Rural Tweets Classification We mapped each ZIP code into its respective urban or rural area according to its Rural-Urban Commuting Area (RUCA) coding [193]. RUCA codes classify US ZIP codes and census tracts into ten levels based on commuting information. For example, Level 1 stands for a major metropolitan area, while level 10 represents an isolated rural area. These levels can be further grouped into four tiers [194, 195]: Urban core (level 1), Suburban (level 2 - 3), Large rural (level 4-6), and Small-town/Rural (level 7 - 10). In this investigation, we focused on the urban core and the small-town/rural, as we anticipated more notable differences would be found at this level.

Preprocessing We removed non-English tweets using the Tweet's lang attribute and the langdetect language detection package (version 1.0.9). For each remaining tweet, we removed URLs, handlers, and the leading "RT". We dropped punctuations and converted all text into lowercase. We then removed tweets with less than three words from the data corpus.

5.3.2 Word Embedding

We trained word embedding models using the skip-gram negative sampling approach implemented in the genism python package (version 4.1.2). We set the vector dimension size to 200 and applied a window size of 5. To learn the monthly sentiment changes of urban and rural users on a monthly basis, we trained word2vec models using the monthly corpus, with ten epochs. For models trained using tweets across months, we went through the corpus five times for efficiency. Parameter tuning was accomplished through word analogy tests. We obtained the word embedding model all-tweets-w2v from all of the tweets. Two separate models, urban-w2v and rural-w2v, were generated using all of the tweets from Urban core and Small-town/Rural respectively.

5.3.3 COVID-19 Hashtag Selection

Twitter users often apply hashtags to label their tweets by topic or theme [196]. Thus, we relied on the hashtags to describe and infer topics about COVID-19. We utilized the word embedding model all-tweets-w2v to find and cluster hashtags related to COVID-19.

The relevance of a hashtag to COVID-19 was measured through a similarity comparison between the given hashtag vector and the vectors for the three most common hashtags in the collected data: #covid19, #covid, and #coronavirus. We defined the relevance score as the maximum of the three cosine similarity values. We selected all hashtags with a relevance score over a certain relevance threshold and a frequency greater than 50 from the all-tweets-w2v model. These hashtags were then subject to an automated clustering process. The relevance threshold is crucial to our analysis. A larger threshold will lead to

a small set of hashtags, resulting in an undersampling of all related hashtags, whereas a smaller threshold will include non-COVID-19 related hashtags. To determine an appropriate relevance threshold, we instructed five human annotators to review hashtags with similarity scores above a threshold and the corresponding clustering quality to judge whether hashtags under the current threshold are related to COVID-19.

Specifically, we relied on two rounds of human evaluation to determine the relevance threshold from a candidate list of [0.4, 0.45, 0.5, and 0.55]. The first round of human evaluation focused on the "recall" or the quantity of collected hashtags, the aim was to find more COVID-19 related hashtags, this round was accomplished by one annotator. For each threshold t, we first defined a score range, [t, t+0.05). Then we asked the annotator to check the 100 randomly selected hashtags with a relevance score in the given score range and label whether the hashtags are related to COVID-19.

The second round of human evaluation focused on the "precision", or the quality of collected hashtags. For each threshold, we collected all relevance hashtags and clustered them into various clusters. Then, we asked five annotators to review the generated clusters, and to judge 1) whether a given cluster is related to COVID-19, and 2) the quality of the given cluster in terms of the similarity of hashtags within the cluster. Each cluster was judged by three annotators independently. The questionnaire is shown in the textbox below.

Each table contains all clusters generated from the given hashtags and the top ten hashtags (sorted by counts) for each cluster. For each cluster, please answer the following two questions with provided dropdown list:

Q1. Based on the hashtags, is this cluster related to covid-19?
Possible Answers: Yes/Maybe/No

If the answer to question 1 is *Yes* or *Maybe*, please continue to answer question 2:

Q2. What is the quality of the cluster? (i.e., are all ten hashtags related to a certain topic? Or are the ten hashtags related to different topics?)
Possible Answers: Good / OK / Bad (corresponding scores: 2/1/0)

5.3.4 Topic Extraction with Hashtag Clustering

We applied UMAP [197] on the vector representation of COVID-19 related hashtags to perform dimensionality reduction and mitigate the impact of an high dimensional system [198]. Clustering was accomplished via HDBSCAN [199]. We performed a grid search on UMAP and HDBSCAN to find the clustering model with the highest relative validity score, a fast approximation of the Density-Based Cluster Validity (DBCV) [200] to evaluate density-based and arbitrarily-shaped clusters. The resulting clusters represent topics related to COVID-19. We define the topic vector as the weighted average of hashtags vectors in the cluster, where the weight is proportional to the count of the hashtag in the corpus. This definition references the general usage of word embedding in document representation [201]. All experiments were performed with the UMAP (version 0.5.2), hdbscan (version 0.8.28) and sklearn (version 1.0.2) python packages.

5.3.5 Opinion Adjectives in SentiWordNet

Opinion adjectives have been adopted to analyze stereotypes through the geometry of word embedding vectors [55, 58]. For example, the vector for the adjective arrogant is close to the vector for men while the vector for elegant is close to the vector for women [55]. For this investigation, we relied on the annotated adjectives in SentiWordNet 3.0 [202] to quantify people's sentiment about COVID-19 topics.

In SentiWordNet 3.0, each adjective, say a, has multiple meanings, and each meaning has a $pos(a)$ and $neg(a)$ score. For example, the word "unable" has three meanings, unable (#1), unable (#2), and unable (#3). The numbers are ordered based on the frequency of use. The most common meaning (#1) is "not having the necessary means or skill or know-how", and the other two meanings (#2, #3) are "lacking necessary physical or mental ability" and "lacking in power or forcefulness". These three meanings have different sentiment scores. For instance, unable (#1) has a positive score of 0.0 and a negative score of 0.75, whereas unable (#2) has scores of 0.0 and 0.375, respectively. In this investigation, we only focused on the $pos(a)$ and $neg(a)$ scores of single SynsetTerm with the most common meaning $a(\#1)$. The $pos(a) + neg(a)$ score distribution for all adjectives in SentiWordNet3.0 are shown in Figure 5.2.

As shown in Figure 5.2, there are two groups of adjectives: the adjectives with $pos(a) + neg(a) \geq 0.5$ and the adjectives with $pos(a) + neg(a) < 0.5$. We considered the adjectives with $pos(a) + neg(a) \geq 0.5$ as "sentiment-rich" adjectives and kept them in the further sentiment analysis.

5.3.6 Sentiment Analysis with Opinion Adjectives

We assumed that adjectives that are more often used to describe a hashtag will have a higher similarity score with respect to the hashtag than those that are infrequently used. In this regard, the difference in the usages of adjectives between urban and rural users can be measured via the difference in the hashtag-adjective similarity scores between urban and

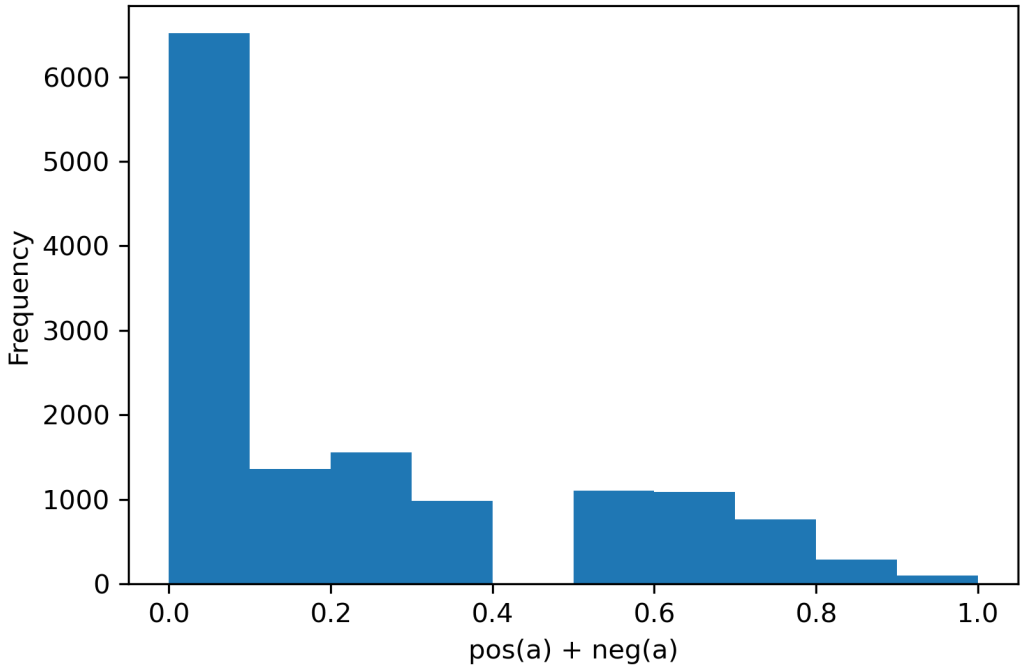

Figure 5.2: The pos(a) + neg(a) score distribution for all adjectives (a) in SentiWordNet3.0.

rural word embedding models. For instance, the adjectives mostly used by urban users to describe a COVID-19 topic can be learned from the comparison between the topic vector to the adjectives in the urban-w2v model. Similarly, the preference of adjectives for rural users can be obtained from rural-w2v model. To ensure a fair comparison between urban and rural areas, we retained adjectives that appeared in both urban-w2v and rural-w2v for sentiment calculation.

We combined the topic-adjective similarity score with the sentiment score for adjectives to learn the sentiment for a topic of interest. Formally, given an adjective collection A, the sentiment score of an adjective a in A, represented as sent (a), is defined as $pos(a) - neg(a)$. The raw sentiment score about a target t in the word2vec model is defined as follows,

$$sent_{raw}(t) = \sum_{a \in A} sim(a,t) \times sent(a) \tag{5.1}$$

where $sim(a,t)$ refers to the cosine similarity between the vector for adjective a and the

vector for target t.

To enable a comparison between two sentiment system, we normalized the raw sentiment score of topics in each model according to their z-score,

$$sent_{nor}(t) = \frac{sent_{raw}(t) - avg(S)}{std(S)} \tag{5.2}$$

where S defines a baseline hashtag set that contains 1000 randomly sampled hashtags. It should be recognized that we normalized urban and rural sentiment score using two different baseline sets, in which hashtags are randomly selected from their own respective vocabularies. The raw sentiment scores for the baseline hashtags were relied upon to estimate the mean $avg(S)$ and standard deviation $std(S)$. The resulting normalized sentiment score reflects the magnitude of positive or negative sentiment, which we apply to compare the differences in urban and rural sentiments.

We utilized a topic vector to represent all of the hashtags in a topic. This approach calculates the sentiment about a topic; however, it cannot estimate the variance across sentiments (i.e., the sentiment difference for various hashtags). Thus, for each topic, we sampled 25% of the hashtags without replacement according to their weights (i.e., proportional to their counts). We then averaged the vectors for these hashtags to obtain a sampled topic vector. The sentiment score for the sampled topic vector is calculated as described earlier. This process was repeated ten times to obtain a set of scores, which were used to compute the average sentiments and their variance.

5.4 Results

5.4.1 Word Embedding Hyperparameter Tuning

The hyperparameters of the word2vec models that were tuned in this study are the vector size, window size, and the number of iterations. These hyperparameters were selected based on a previous study of word2vec [203]. We used one month of collected tweets (2021-05) as the text corpus to train different word embedding models by using grid

search. Then, a word analogy test was performed to select the appropriate parameters for further word embedding model training.

For the word analogy test, we followed the work of Mikolov et al. [40] The test accuracy of different parameter settings are shown in Table 5.1. To obtain a more intuitive comparison of the 14 tests, we depicted the accuracy change in Figure 5.3. In Figure 5.3, the first row in x-axis shows the window size, the second row shows the number of trained iterations, the third raw shows the selected vector size. The y-axis represents the analogy test accuracy.

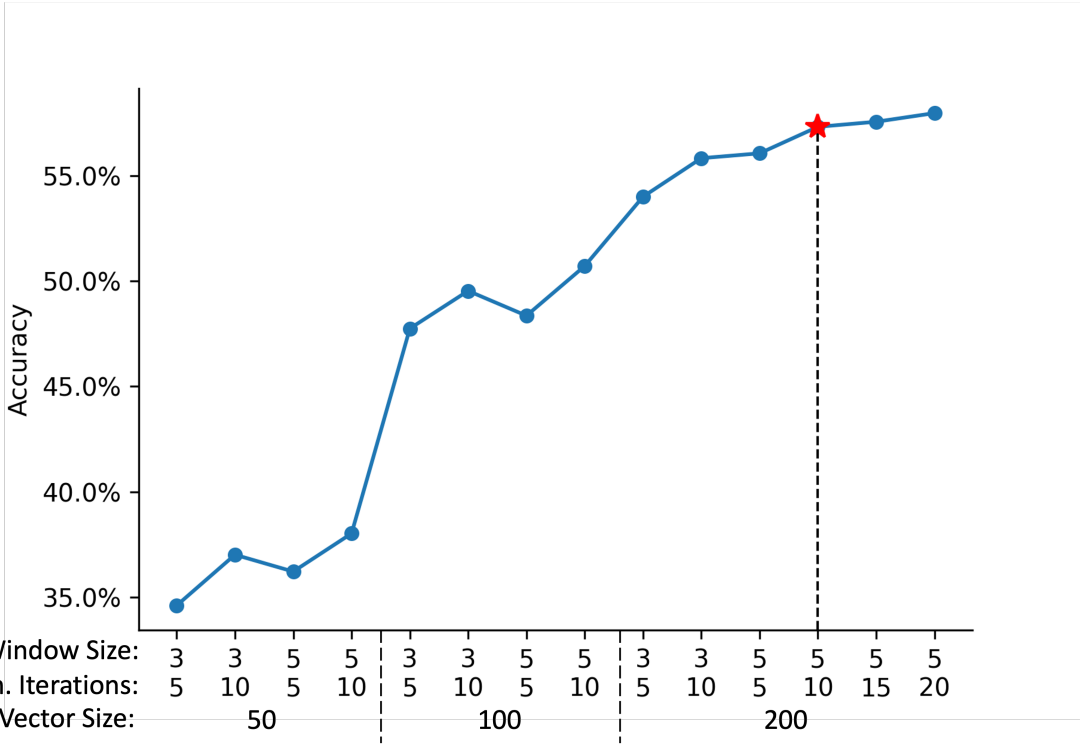

Figure 5.3: Word analogy test accuracy for 14 different parameter settings.

As shown in Figure 5.3, increasing vector size had the most remarkable improvement in test accuracy among the three candidate hyperparameter values. Thus, the vector dimension was set to 200, a common choice in Twitter word embedding training [43]. As for window size, we limited the maximum window size to 5. This is because the average number of words per tweet in our collected data was 10.44; setting a window size to greater

than 5 implies that the vector of the current word depends on a word from the words in another tweet, which is unreasonable. Based on experiments result, we set the window size to 5. Finally, a comparison of experiments with vector size of 200 (the rightmost six dots in Figure 5.3) showed that training the corpus more than ten times had limited improvements, the gain of the accuracy was quite small after the hyperparameters of (5, 10, 200). Considering the model training time, we set the number of iterations to 10 for the monthly corpus.

Table 5.1: Results of word analogy tests.

Experiment ID	Vector size	Window size	Num. of iterations	Accuracy
1	50	3	5	34.60%
2	50	3	10	37.01%
3	50	5	5	36.21%
4	50	5	10	38.02%
5	100	3	5	47.75%
6	100	3	10	49.54%
7	100	5	5	48.36%
8	100	5	10	50.72%
9	200	3	5	54.01%
10	200	3	10	55.84%
11	200	5	5	56.07%
12	200	5	10	57.33%
13	200	5	15	57.57%
14	200	5	20	57.98%

5.4.2 Urban and Rural Tweets Distribution

Figure 5.4 depicts the number of tweets collected in the US, where blue represents Urban core areas and red represents small-town/rural areas. A darker color means that there were a higher number of tweets in that area. As can be seen, the figure generally matches the urban rural classification scheme in the US [204]. Table 5.2 provides summary statistics for the three word embedding models trained using all of the tweets.

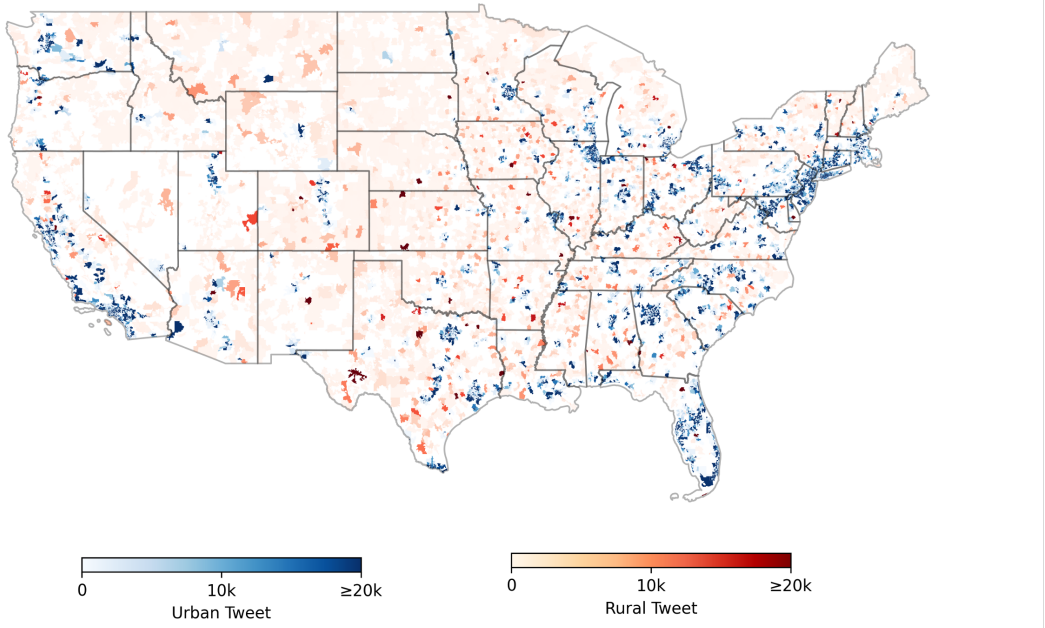

Figure 5.4: Number of tweets collected in US urban core and small-town/rural ZIP codes.

5.4.3 Human Evaluation Results of Hashtag Relevance Thresholds

As discussed in the Method section, we relied on human evaluation to determine the relevance threshold from a candidate list of [0.4, 0.45, 0.5, 0.55]. The result of first round human evaluation, the "recall" round, are shown in Table 5.3.

From Table 5.3, it can be seen that there are still some COVID-19 relevant hashtags below threshold 0.55, setting a threshold equal to or higher than 0.55 will result in an incomplete collection of COVID-19 relevant hashtags. Therefore, in the second round of human evaluation, we removed 0.55 from the threshold candidates.

Our second round of human evaluation focused on the quality of hashtag clusters. For clusters related to COVID-19 (clusters with an answer yes or maybe for question 1), we report reviewers' Kappa agreement on question 1 and averaged cluster quality score (question 2) in Table 5.4.

Since threshold 0.5 resulted in the highest kappa agreement score and the highest quality score, we selected 0.5 as the relevance threshold for the further analysis.

Table 5.2: Training data for the word embedding models. (RUCA: Rural-Urban Commuting Area).

RUCA Tier	**All Tweets**	**Urban Core**	**Small-town/Rural**
Number of tweets	407 million	350 million	18 million
Words per tweet	10.47	10.44	10.54
Hashtags per tweet	0.18	0.18	0.17
Word2vec model	*all-tweets-w2v*	*urban-w2v*	*rural-w2v*

Table 5.3: Number of COVID-19 related hashtags in 100 random samples at different threshold level, labeled by one annotator.

ID	Threshold	Score range	COVID-19 hashtags	Example hashtags
1	0.55	[0.55, 0.59)	> 70	faucihero, unmaskamerica
2	0.5	[0.50, 0.55)	$40 \sim 50$	firefauci, quarentinelife
3	0.45	[0.45, 0.50)	$10 \sim 20$	maskupnola, backtoschool2020
4	0.4	[0.40, 0.45)	< 5	coronavirussicilia, backtobusiness

5.4.4 Topic Clustering

We collected 2,666 COVID-19 related hashtags, which clustered into 30 distinct topics. After a manual review of the clusters, we found that 20 of the topics were closely related to COVID-19 in US. The other 10 less relevant topics included those related to social justice (e.g., the George Floyd events), news about the Middle East and COVID-19 in other countries (e.g., Canada, India, and Mexico). Figure 5.5 presents a 2D representation of the word embedding vectors for the clustered hashtags in 20 COVID-19 related topics. Based on closeness of the topic hashtags, we further grouped the topics into four categories: Misinformation, Prevention and Treatment, Economy, and News and Politics. For example, topics belonging to the Misinformation category, including, Covidiots, China virus, and plandemic, all appear in the upper left corner. Topics about News and Politics are grouped

Table 5.4: Human evaluation of hashtag cluster quality.

Id	Threshold	Fleiss' kappa of Q1 on COVID-19 related clusters	Average clusters' quality score (Q2)
1	**0.50**	**0.364**	**1.545**
2	0.45	0.253	1.487
3	0.40	0.340	1.483

together in the upper right corner. The topics of Prevention and Treatment and Economy also exhibit a similar grouping pattern. Certain topics, namely COVID-19, Health, and School, do not fall into the four categories.

Table 5.5 shows the number of hashtags, the 10 most tweeted hashtags and a manually assigned label for each of the 20 topics. The topics are presented in descending order according to the number of unique hashtags they hold. The hashtags are presented in descending order according to their frequency. It can be seen that the Mandates, Health and Vaccine topics are affiliated with the most user-generated hashtags, which highlights the users' concerns about COVID-19 prevention and its impact on health.

Figure 5.6 shows the trends in volume and relevance for COVID-19 for the selected topics and categories. The black line in Figure 5.6 indicates the number of monthly new COVID-19 cases in the US, where EUA stands for emergency use authorization. Specifically, Figure 5.6A shows the volume of tweets for a topic (number of tweets that contain at least one of the topic hashtags), while Figure 5.6B shows the relevance of topics to COVID-19. There are several notable observations that are worth highlighting here. First, there is an overall declining trend in COVID-19 related tweets. Secondly, the most tweeted topic changes overtime. For instance, prior to February 2021, the most tweeted topic was Mandates. Afterwards, Vaccines became the most tweeted, as well as the most relevant topic, with monthly discussions peaking in April 2021. This trend is positively correlated with the changes in the number of vaccinated people in the US. Third, for topics in the News and

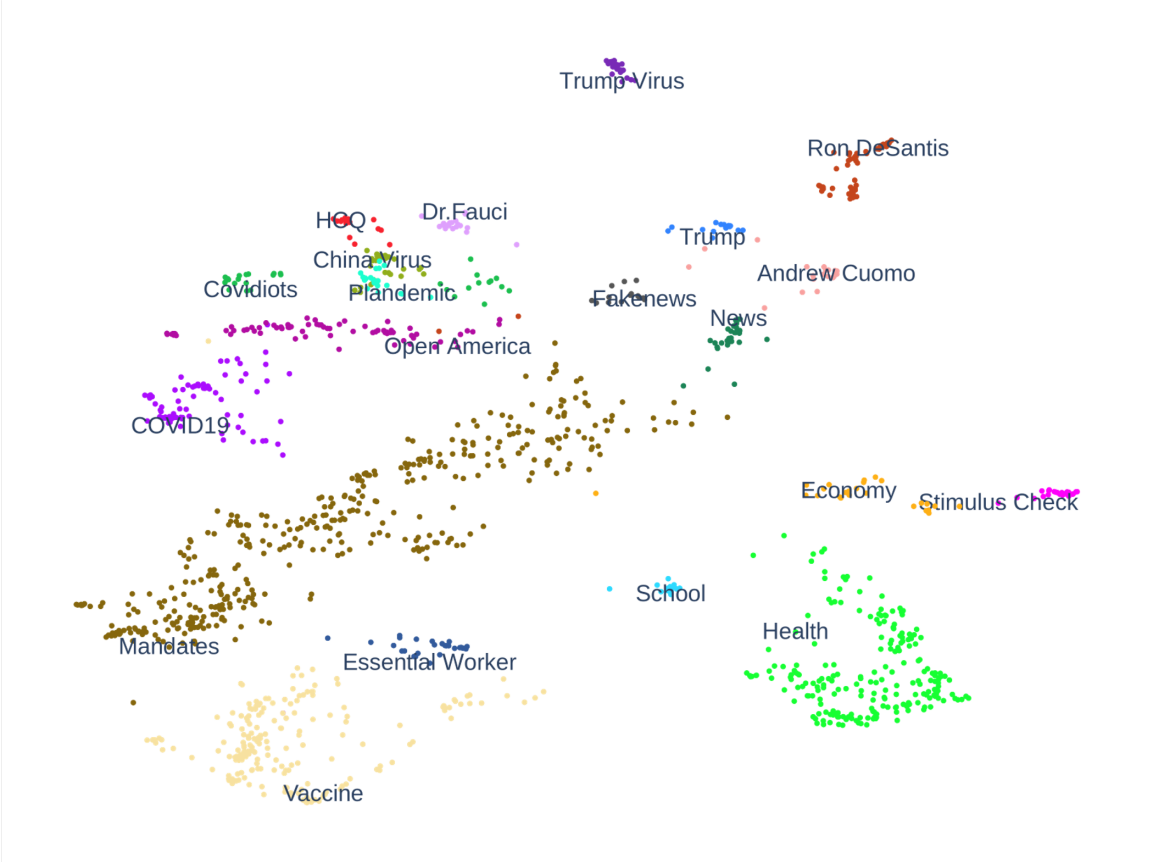

Figure 5.5: A 2D representation of UMAP clustering results of 20 topics. Each point represents a distinct hashtag.

Politics category, we found that the changes in topic Trump, both in volume and COVID-19 relevance, are aligned with progress in the 2020 presidential election. The relevance of the Donald Trump topic to COVID-19 reached its highest level in October 2020, after which it gradually diminished. Finally, we observed that before 2022, the trend of Misinformation category generally matched with the change in the number of COVID-19 new cases. Yet, after 2022, there was a decline in both the volume and relevance scores across all but one topic (Vaccine), although the number of new COVID-19 cases peaked in January.

5.4.5 Urban versus Rural Sentiment

Figure 5.7 depicts the normalized urban and rural sentiments about COVID-19 related topics. In Figure 5.7, the category ID for each topic is shown to the left of the topic name.

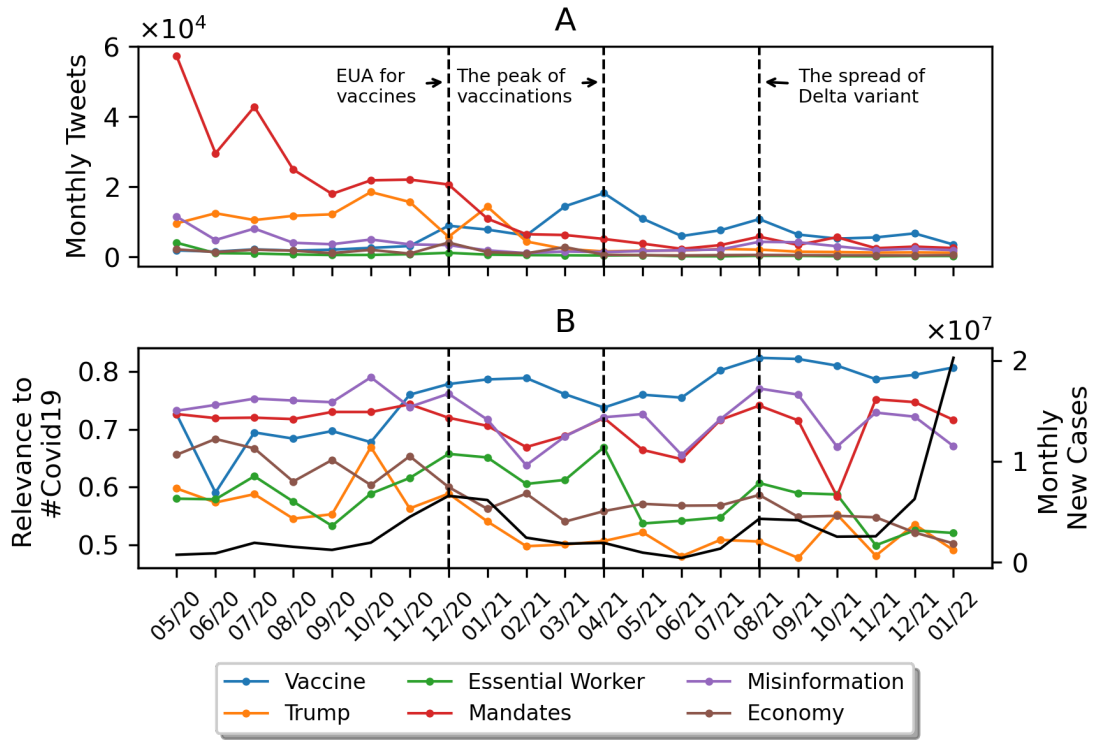

Figure 5.6: The monthly trend in volume (A) and relevance to COVID-19 (B) for selected topics and categories.

The error bar indicates $+/-$ one standard deviation of the sentiment. The three additional topics at the bottom (separated by the dotted lines) are displayed to provide readers with some intuition into the degree of positivity (negativity) represented by the sentiment score. The raw P-values of Welch's t-tests are shown in the right column, where the bold text indicates a statistically significant difference ($P < .05/20$) after Bonferroni correction. We normalized urban and rural scores using their baseline hashtag set. For urban-w2v, the mean raw sentiment score (prior to normalization) was -4.58 with a standard deviation of 5.84. For rural-w2v, the mean raw sentiment score was -11.02, with a standard deviation of 7.43.

As shown in Figure 5.7, both urban and rural users exhibited negative sentiments for the majority of COVID-19 related topics. The only topic with a positive sentiment was Essential Worker. Both urban and rural users communicated weak negative sentiments

Table 5.5: The 20 COVID-19 topics inferred from collected tweets.

Category	Topic Label	10 Most Frequent Hashtags	Unique Hashtags
	COVID19	covid19, coronavirus, covid, covid_19, pandemic, covid-19, corona, covid__19, omicron, covid–19	79
#1 Misinformation	Open America	europenamericanow, nomasks, vaccinemandate, maskmandate, nomask, donotcomply, reopenamerica, vaccinepassport, vaccinepassports, maskmandates	73
	Covidiots	covidiots, antivaxxers, idiots, moron, covididiots, stupidity, morons, antimaskers, antivaxxer, antivax	30
	China Virus	chinavirus, billgates, ccpvirus, wuhanvirus, wuhan, chinaliedpeopledied, chinesevirus, chinaliedandpeopledied, agenda21, wuhancoronavirus	30
	Dr. Fauci	fauci, drfauci, firefauci, faucithefraud, anthonyfauci, fauciliedpeopledied, fauciemails, faucilied, faucifraud, birx	22
	Plandemic	plandemic, hoax, scamdemic, factsnotfear, covidhoax, fearmongering, kungflu, scamdemic2020, fearporn, coronahoax	20
	HCQ	hydroxychloroquine, ivermectin, cnntownhall, remdesivir, hcq, regeneron, hydroxycloroquine, trumpvaccine, hydroxycholoroquine, dexamethasone	16
#2 Prevention and Treatment	Mandates	wearamask, 2020, staysafe, maskup, stayhome, socialdistancing, quarantine, quarantinelife, mask, lockdown	397
	Vaccine	covidvaccine, vaccine, science, getvaccinated, vaccinated, pfizer, moderna, getvaccinatednow, vaccineswork, covid19vaccine	198
	Essential Worker	essentialworkers, nurses, healthcareheroes, inthistogether, healthcareworkers, frontlineworkers, frontlineheroes, healthcareworker, frontliners, frontlines	27

#3 Economy	Stimulus Check	stimuluscheck, stimulus, unemployment, heroesact, americanrescueplan, stimuluspackage, covidrelief, caresact, covidreliefbill, stimulusbill	28
	Economy	economy, housing, homelessness, unemployed, markets, debt, economic, evictionmoratorium, jobsreport, housingcrisis	26
#4 News and Politics	Ron DeSantis	deathsantis, desantis, rondesantis, gregabbott, deathdesantis, desantisfailedflorida, floridacovidepicenter, harriscounty, floriduh, desantisvariant	58
	Trump Virus	trumpvirus, trumpknew, trumpliesamericansdie, trumpfailedamerica, trumphasnoplan, trumpliedpeopledied, trumpisanidiot, trumpownseverydeath, trumpgate, trumpliespeopledie	31
	News	foxnews, news, cnn, breakingnews, journalism, nytimes, abcnews, nyt, newyorktimes, nbcnews	31
	Andrew Cuomo	cuomo, deblasio, killercuomo, andrewcuomo, governor, chriscuomo, fredo, cuomokilledgrandma, cuomocoverup, governorcuomo	21
	Donald Trump	trump, donaldtrump, potus, whitehouse, realdonaldtrump, presidenttrump, pence, mikepence, potus45, donaldtrumpjr	14
	Fake News	fakenews, lies, factcheck, propaganda, misinformation, conspiracytheory, disinformation, mainstreammedia, factchecking, bantiktok	14
	Health	health, cancer, anxiety, depression, publichealth, hiv, diabetes, medicine, doctor, breastcancer	219
	Reopen School	schools, schoolsreopening, schoolreopening, lausd, stayinformed, reopeningschools, nycdoe, publicschools, virtualuntilsafe, dpa	17

(in [-1, 0]) for the Mandates, Vaccine and Health topics. By contrast, there was a strong negative sentiment (around -2) for topics about News, Politics, and Misinformation for both groups.

For topics related to COVID-19 prevention (i.e., Vaccine and Mandates), it was ob-

served that rural users exhibited a stronger negative sentiment than urban users. For topics related to misinformation and conspiracy theory, it was observed that urban users expressed a much stronger negative feeling about the Covidiots and Fake news topics, while rural users tended to use adjectives with stronger negative sentiment when discussing Open America, plandemic, and Dr. Fauci. For Politics related topics, we observed a clear political divide when comparing the urban and rural users for their sentiment toward political figures. Urban users talked about Donald Trump and Ron DeSantis - two Republicans - with stronger negative sentiment, while rural users were more likely to criticize Andrew Cuomo - a Democrat. The urban vs. rural sentiment differences in prevention and politics related topics are statistically significant ($P < .001$ with Bonferroni correction).

To gain further intuition into the degree of positivity (negativity) represented by the sentiment score, Figure 5.7 also includes three additional topics for comparison: Christmas, Thanksgiving and Election 2020. Among the three topics, Christmas and Thanksgiving have positive sentiments, ranging from 0.5 to 0.8, whereas the topic Election 2020 has a negative sentiment of around -0.5.

5.4.6 Topic Sentiment Temporal Trend

The temporal trend with respect to topic sentiment was characterized as change in monthly sentiment change. However, some topics and their hashtags only appeared in a certain month. For example, in our collected rural tweets, the hashtags about Reopen School topic only appeared in July and August 2020, which indicates that there are no monthly sentiment changes for this topic. As a result, we removed the 11 topics with insufficient hashtags or similar urban and rural sentiment trends here. Figure 5.8 depicts the trend in the monthly sentiment for the nine remaining topics. The monthly sentiment for all 20 topics was presented in Figure 5.9.

In Figure 5.8, each month was depicted on the x-axis. The center of a dot represents the sentiment value of the topic, where the size of the dot reflects the ratio of the volume of

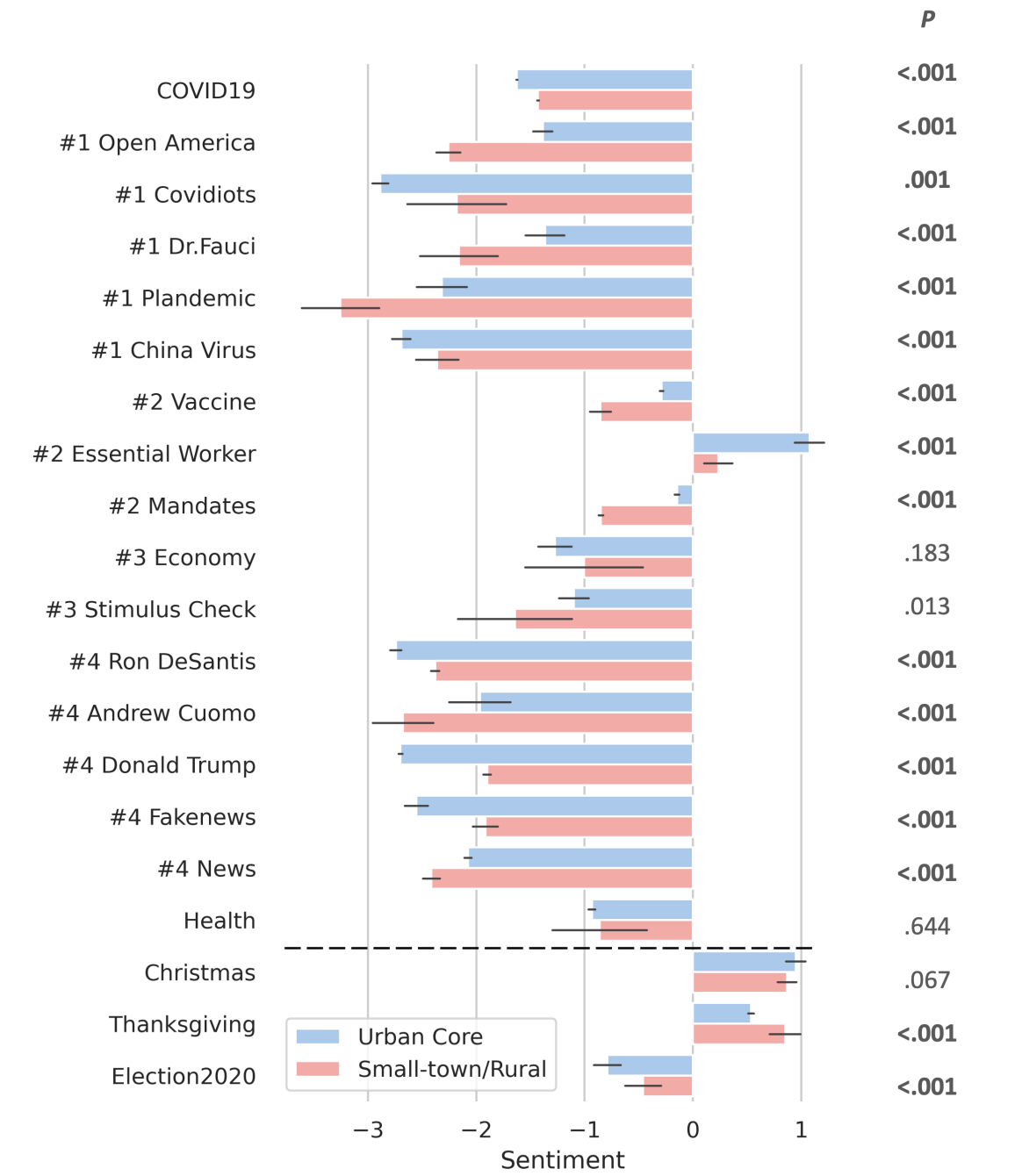

Figure 5.7: Overall normalized urban and rural sentiment towards COVID-19 and selected 20 topics.

topic's current month's tweets to the sum of the topic's tweets for all months. The trendlines correspond to a locally weighted linear regression for Urban Core and Small-town/Rural.

As shown in Figure 5.8, in general, both urban and rural Twitter users' attitudes about COVID-19 gradually became more negative. One of the exceptions, shown in the first row of Figure 5.8, was the Economy topic, for which urban users appeared to transition from a negative to a positive sentiment. A possible reason for this change is that the US economic recovery initiated in late 2021 [205]. The second row of Figure 5.8 shows the sentiment trend about three celebrities. For the topic about Dr.Fauci, December 2020 was a watershed moment in the public's attitudes about him, when he accepted the offer to become the chief medical advisor to the president in the Biden administration. For politicians, Rural users' sentiment towards Donald Trump and Ron DeSantis is consistently higher than that of urban users. The temporal trends for sentiment about prevention-related topics are depicted in the third row of Figure 5.8, where urban and rural users show a similar, gradually declining trend towards Vaccine and Mandates. While rural users had a relatively stable sentiment towards the topic of Essential workers, urban users' sentiment slowly became negative.

5.5 Discussion

5.5.1 Principle Findings

We believe that the approach for learning public sentiment introduced in this chapter has several benefits over prior methods. First, by combining the word embedding models with sentiment-rich opinion adjective lexicons, users of this approach can conduct sentiment analysis in the learned semantic vector space. This allows users to directly infer the sentiment of a population group towards a topic. In comparison to tweet-level sentiment analysis, one advantage of this approach is that it does not require to identify COVID-19 related tweets by using either keyword filters or machine learning classifiers, which is more robust to the noise (e.g., misspellings, synonyms and abbreviations) within the online data.

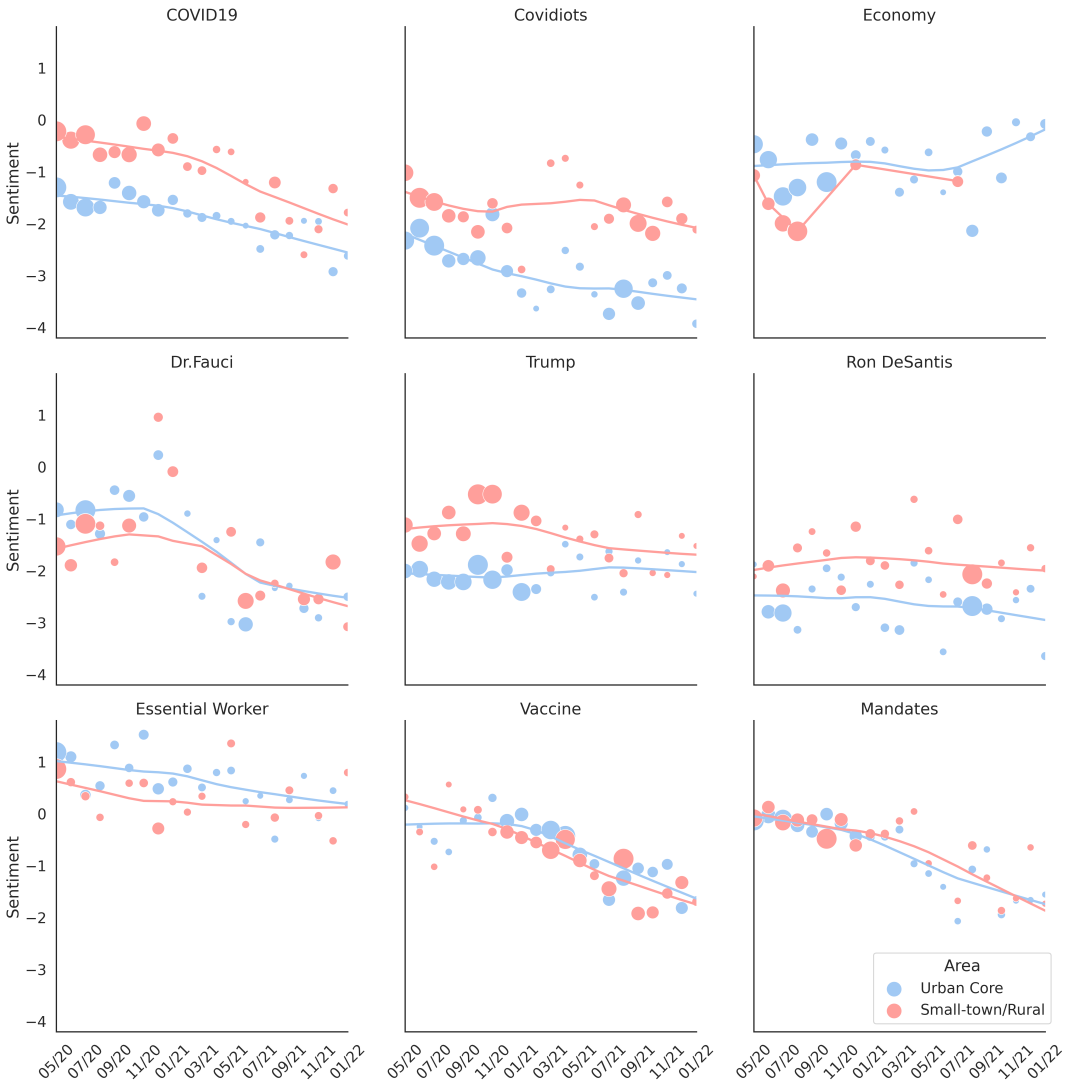

Figure 5.8: Monthly urban and rural sentiment regarding COVID-19 related topics.

Second, unlike commonly used topic modeling techniques, such as LDA, the new method utilizes word embedding vector clustering to identify hashtags and topics of public interest, which works well on tweets, the large amount of noisy short text data. Third, while our approach was tailored to applied in the sentiment analysis of COVID-19, we believe that the trained word embedding models can be directly used for sentiment analysis of other social events, without the hassle of a new round of data collection and labeling. For instance, our data collection period covers the time of the 2020 presidential election, thus the trained model can be directly used for election-related sentiment analysis. Another possible application of this model would be to build a topic extraction and sentiment analysis platform, where users can input any event of interest to obtain related topics and to infer public's sentiments about this event in rural or urban areas. Our learned word embedding models are publicly available on GitHub [206].

At the same time, there are several notable findings from this investigation. First, we observed that urban and rural users clearly harbor different sentiment about certain COVID-19 related topics. In particular, urban users exhibited a stronger negative sentiment about Covidiots, China Virus, Economy and Fake news. By contrast, rural users show a stronger negative sentiment towards Plandemic, Dr.Fauci, and prevention strategies (Vaccine and Mandates). These findings, are consistent with those of prior investigations [8, 182]. Callaghan and colleagues [8] found that rural residents are less likely to "participate in several COVID-19-related preventive health behaviors", Chauhan and colleagues [182] observed that rural residents are less concerned about the coronavirus. Moreover, we observed a clear political divide between urban and rural users through the sentiment analysis of three politicians. For instance, during the time window covered of this study, urban users viewed Andrew Cuomo more favorably than Donald Trump and Ron DeSantis, while vice versa could be said for rural users. These findings are also consistent with studies on political polarization [207]. All of these provide evidence that, with our proposed model, social media data can be effectively leveraged to gain timely insight into the public

understanding and/or sentiment to hot social events.

5.5.2 Limitations

There are also several limitations to this study, which we believe serve as opportunities for future research. First, we relied on tweet's place attribute to obtain user's geolocation, and to infer user's urban/rural information. This step is not 100% accurate, as there are several "non-formatted" places in the collected tweets. The "non-formatted" place can be ambiguous, such as "McDonalds", or too general, such as "Iowa, USA". Through manual review of 200 random sampled tweets, we found 19 (10%) tweets with "non-formatted" place attribute. The geocoding result of the "non-formatted" place attribute may make our result less significant than the true urban vs. rural difference. Second, to quantify the sentiment of a particular group, our method requires training a word embedding model for that group. Our method is less effective if the goal is to compare multiple social groups with different demographics. The issue may be resolved with the assistance of word embedding geometry [208]: performing sentiment analysis on the subspace of the aspect of interest.

5.5.3 Conclusion

In this study, we introduced a novel approach to characterize the public's sentiment about COVID-19 and related topics. By applying topic recognition and a subsequent sentiment analysis, we discovered a clear difference between urban and rural users in their sentiment about COVID-19 prevention strategies, misinformation, politicians and the economy. While these findings might not be representative of the sentiment of the American public more broadly, we believe that such investigations could help policymakers obtain a more comprehensive understanding of the sentiment differences between urban and rural areas on COVID-19 and related topics, so that more targeted deployment of epidemic prevention efforts can be made. Finally, we wish to highlight that our approach is not limited to COVID-19 and it can readily be extended to other topics of interest without additional data collection or model training.

Figure 5.9: Monthly urban and rural sentiment regarding 20 COVID-19 related topics.

CHAPTER 6

Conclusion

6.1 Summary

This dissertation investigates how to utilize user-generated data to understand public perception of societal concerns. Our investigation revolves around two core questions: 1) Given the societal concern of interest, what are the topics of public concern, and 2) what is the public sentiment about the societal concern and related topics. These questions are answered through three related, but computational distinct, tasks.

First, we investigate the behavior of online research cohort membership disclosure. We gathered and analyzed 4,020 tweets and uncovered over 100 tweets disclosing the individuals' memberships in over 15 medical research programs. Through sentiment analysis, we learned that 45.3% of self-disclosed users have a positive attitude towards the joined research project. Our investigation showed that self-disclosure on social media can reveal participants' membership in research cohorts, and such activity might lead to the leakage of a person's identity, genomic, and other sensitive health information.

In the second task, to gain a deeper understanding of the self-disclosure behavior, we investigated the face image sharing trend in a Direct-to-Consumer genetic testing forum. Through topic modeling and statistical inference on over 15,000 Reddit posts, we found that posts including a face received 60% more comments and had karma scores (upvotes – downvotes) 2.4 times higher than other posts. The topics in posts including a face were primarily about sharing, discussing ancestry composition, or sharing family reunion photos. The association between face image posting and a greater level of attention suggests that people are forgoing their privacy in exchange for attention from others. To mitigate this risk, platform organizers and moderators could inform users about the risk of posting face images in a direct, explicit manner to make it clear that their privacy may be compromised

if personal images are shared.

In the final part of this dissertation, we focus on a topic that affects the general population: public sentiment about COVID-19 and related topics. In this investigation, we combined word embedding models with clustering strategies to identify topics closely related to COVID-19, and relied upon the similarity between topic hashtags and opinion adjectives to infer the sentiment with respect to the identified topics. We discovered a significant difference between US urban and rural users in their sentiment about COVID-19 prevention strategies, misinformation, politicians, and the economy. This investigation provides a more comprehensive assessment of the differences in sentiment between urban and rural areas, which, provide intuition into the challenges of geographically-targeted deployment of epidemic prevention and management efforts. The sentiment analysis approach is notable in that it can readily be extended to other topics of interest without additional data collection or model training.

6.2 Future Investigations

Despite the findings in this investigation, there are certain limitations to this work, which pose as next steps for research.

First, to quantify the sentiment of a particular group, our word-embedding based sentiment analysis approach requires training a word embedding model for that group. This method is less effective if the goal is to compare multiple social groups with different demographics. For example, if the user want to compare the sentiment difference between male and female Twitter users, they need to train two separate word embedding models, one with tweets posted by male and the other one with tweets posted by female. The issue may be resolved with the assistance of word embedding geometry[208], by finding a set of male-female concept pairs (e.g., boy-girl, man-women) and project the adjective vectors to the subspace consists of the vectors of male-female concept pairs. This approach can be extended to support the sentiment analysis of a combination of different social groups. For

instance, if the user is interested in the sentiment difference between urban male group and rural female group, we can do so by finding two group of concept pairs: male-female and urban-rural, and then project the word embedding system to the subspace consists of the two groups of vector pairs.

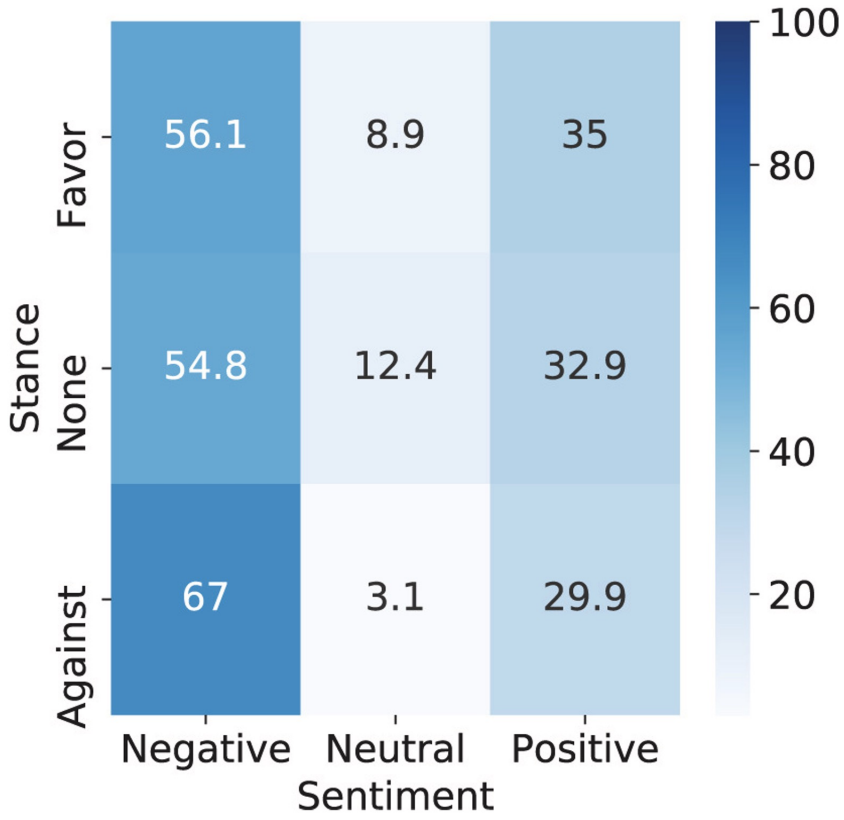

Figure 6.1: Stance and sentiment distribution of SemEval 2016 Task 6a dataset, image from the work of ALDayel and Magdy [1].

Second, our investigation relied upon sentiment analysis to gain intuition into public perception. However, it should be noted that the result of sentiment analysis may differ from the public perception occasionally. This is because people express their opinion toward a target using various tones with different sentiments. For example, one may use either a straight-forward negative sentiment, a conservative neutral sentiment, or even a sarcastic positive sentiment to express an opposing stance towards a target. This entanglement between stance and sentiment makes it harder to learn the true opinion. Figure 6.1 below, from the work of ALDayel and Magdy [1], illustrates the complicated relationship

between sentiment and stance. Error analysis of stance detection models [1, 93] showed that the performance of models often suffers from sentiment entangles. The model can be disguised by the sentence's sentiment and generates a wrong stance.

There are several research directions on uncovering perception from sentences using sentiment information. Disentangle learning [209] can be utilized to remove sentiment distraction from stance detection. While studies in multi-task learning (MTL) [210–212] have shown that training related tasks, such as sentiment and stance detection, together can have better performance than separate, single-task training.

References

[1] Abeer ALDayel and Walid Magdy. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597, 2021.

[2] David J. Ball and Sonja Boehmer-Christiansen. Societal concerns and risk decisions. *Journal of Hazardous Materials*, 144(1):556–563, 2007.

[3] James W. Hazel, Catherine Hammack-Aviran, Kathleen M. Brelsford, Bradley A. Malin, Laura M. Beskow, and Ellen Wright Clayton. Direct-to-consumer genetic testing: Prospective users' attitudes toward information about ancestry and biological relationships. *PLOS ONE*, 16(11):1–20, 11 2021.

[4] Grayson L Ruhl, James W Hazel, Ellen Wright Clayton, and Bradley A Malin. Public attitudes toward direct to consumer genetic testing. In *AMIA Annual Symposium Proceedings*, volume 2019, page 774. American Medical Informatics Association, 2019.

[5] Brooke Auxier, Lee Rainie, Monica Anderson, Andrew Perrin, Madhu Kumar, and Erica Turner. Americans and privacy: Concerned, confused and feeling lack of control over their personal information. *Pew Research Center*, Novemeber 2019. https://tinyurl.com/59y83ect.

[6] Nathan F. Dieckmann, Robin Gregory, Terre Satterfield, Marcus Mayorga, and Paul Slovic. Characterizing public perceptions of social and cultural impacts in policy decisions. *Proceedings of the National Academy of Sciences*, 118(24):e2020491118, 2021.

[7] Tristan McCaughey, David M Budden, Paul G Sanfilippo, George EC Gooden, Li Fan, Eva Fenwick, Gwyneth Rees, Casimir MacGregor, Lei Si, and Christine Chen. A need for better understanding is the major determinant for public perceptions of human gene editing. *Human gene therapy*, 30(1):36–43, 2019.

[8] Timothy Callaghan, Jennifer A Lueck, Kristin Lunz Trujillo, and Alva O Ferdinand. Rural and urban differences in covid-19 prevention behaviors. *The Journal of Rural Health*, 37(2):287–295, 2021.

[9] Amanda Barna. Why is public opinion research important? *CMOR*, 2011. https://www.cmoresearch.com/articles/why-is-public-opinion-research-important.php.

[10] Ralph Nafziger. Em 4: are opinion polls useful? *Historians.org*, January 1946. https://www.historians.org/about-aha-and-membership/aha-history-and-archives/gi-roundtable-series/pamphlets/em-4-are-opinion-polls-useful-(1946).

[11] Paul Burstein. The impact of public opinion on public policy: A review and an agenda. *Political research quarterly*, 56(1):29–40, 2003.

[12] Xuehua Han and Juanle Wang. Using social media to mine and analyze public sentiment during a disaster: A case study of the 2018 shouguang city flood in china. *ISPRS International Journal of Geo-Information*, 8(4):185, 2019.

[13] Zheye Wang and Xinyue Ye. Social media analytics for natural disaster management. *International Journal of Geographical Information Science*, 32(1):49–72, 2018.

[14] Martin Gilens and Benjamin I Page. Testing theories of american politics: Elites, interest groups, and average citizens. *Perspectives on politics*, 12(3):564–581, 2014.

[15] Zhijun Yin, Bradley Malin, Jeremy Warner, Pei-Yun Hsueh, and Ching-Hua Chen. The power of the patient voice: learning indicators of treatment adherence from an online breast cancer forum. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[16] Zhijun Yin, Lijun Song, Ellen W Clayton, and Bradley A Malin. Health and kinship matter: Learning about direct-to-consumer genetic testing user experiences via online discussions. *PLOS ONE*, 15(9):e0238644, 2020.

[17] Salman Aslam. Twitter by the numbers (2022): stats, demographics & fun facts_2022. *Omnicore Agency*, February 2022. https://www.omnicoreagency.com/twitter-statistics/.

[18] Georgi Todorov. 70+ important reddit statistics 2022. *Learn Digital Marketing*, December 2021. https://thrivemyway.com/reddit-statistics/.

[19] Zhijun Yin, Lina M Sulieman, and Bradley 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, 2019.

[20] Yang Liu and Zhijun Yin. Understanding weight loss via online discussions: content analysis of reddit posts using topic modeling and word clustering techniques. *Journal of Medical Internet Research*, 22(6):e13745, 2020.

[21] Ari Z Klein, Arjun Magge, Karen O'Connor, Jesus Ivan Flores Amaro, Davy Weissenbacher, and Graciela Gonzalez Hernandez. Toward using twitter for tracking covid-19: a natural language processing pipeline and exploratory data set. *Journal of Medical Internet Research*, 23(1):e25314, 2021.

[22] Yongtai Liu, Zhiyu Wan, Weiyi Xia, Murat Kantarcioglu, Yevgeniy Vorobeychik, Ellen Wright Clayton, Abel Kho, David Carrell, and Bradley A Malin. Detecting the presence of an individual in phenotypic summary data. In *AMIA Annual Symposium Proceedings*, volume 2018, page 760. American Medical Informatics Association, 2018.

[23] Yongtai Liu, Chao Yan, Zhijun Yin, Zhiyu Wan, Weiyi Xia, Murat Kantarcioglu, Yevgeniy Vorobeychik, Ellen Wright Clayton, and Bradley A Malin. Biomedical research cohort membership disclosure on social media. In *AMIA Annual Symposium*

Proceedings, volume 2019, page 607. American Medical Informatics Association, 2019.

[24] Yongtai Liu, Zhijun Yin, Zhiyu Wan, Chao Yan, Weiyi Xia, Congning Ni, Ellen Wright Clayton, Yevgeniy Vorobeychik, Murat Kantarcioglu, and Bradley A Malin. Implicit incentives among reddit users to prioritize attention over privacy and reveal their faces when discussing direct-to-consumer genetic test results: Topic and attention analysis. *JMIR Infodemiology*, 2(2):e35702, 2022.

[25] Prafulla Bafna, Dhanya Pramod, and Anagha Vaidya. Document clustering: Tf-idf approach. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 61–66. IEEE, 2016.

[26] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[27] Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1):1–22, 2016.

[28] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[29] Daniel Ramage, Evan Rosen, Jason Chuang, Christopher D Manning, and Daniel A McFarland. Topic modeling for the social sciences. In *NIPS 2009 workshop on applications for topic models: text and beyond*, volume 5, pages 1–4, 2009.

[30] Hae-Wol Cho. Topic modeling. *Osong public health and research perspectives*, 10(3):115, 2019.

[31] David Blei and John Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

[32] Ike Vayansky and Sathish AP Kumar. A review of topic modeling methods. *Information Systems*, 94:101582, 2020.

[33] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456, 2013.

[34] Margaret E Roberts, Brandon M Stewart, Dustin Tingley, and Edoardo M Airoldi. The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*, volume 4, pages 1–20. Harrahs and Harveys, Lake Tahoe, 2013.

[35] Theo Lebryk. Introduction to the structural topic model (stm). *Medium*, Apr 2021. https://towardsdatascience.com/introduction-to-the-structural-topic-model-stm-34ec4bd5383.

[36] Margaret E Roberts, Brandon M Stewart, and Dustin Tingley. Stm: An R package for structural topic models. *Journal of Statistical Software*, 91:1–40, 2019.

[37] Congning Ni, Zhiyu Wan, Chao Yan, Yongtai Liu, Ellen Wright Clayton, Bradley Malin, and Zhijun Yin. The public perception of the# geneeditedbabies event across multiple social media platforms: Observational study. *Journal of Medical Internet Research*, 24(3):e31687, 2022.

[38] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394, 2010.

[39] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247, 2014.

[40] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[41] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *ArXiv*, 2013. https://arxiv.org/abs/1301.3781.

[42] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[43] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[44] Siwei Lai, Kang Liu, Shizhu He, and Jun Zhao. How to generate a good word embedding. *IEEE Intelligent Systems*, 31(6):5–14, 2016.

[45] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the association for computational linguistics*, 3:211–225, 2015.

[46] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1555–1565. Association for Computational Linguistics, June 2014.

[47] Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327. Association for Computational Linguistics, June 2013.

[48] Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C.-C. Jay Kuo. Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, 8:e19, 2019.

[49] Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35. Association for Computational Linguistics, August 2016.

[50] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

[51] Yang Li and Tao Yang. Word embedding for understanding natural language: A survey. *Guide to Big Data Applications*, pages 83–104, 2018.

[52] Billy Chiu, Anna Korhonen, and Sampo Pyysalo. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 1–6. Association for Computational Linguistics, August 2016.

[53] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150. Association for Computational Linguistics, June 2011.

[54] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, nov 2011.

[55] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

[56] S. Selva Birunda and R. Kanniga Devi. A review on word embedding techniques for text classification. In *Innovative Data Communication Technologies and Application*, pages 267–281. Springer Singapore, 2021.

[57] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[58] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[59] Caroline Criado Perez. *Invisible women: Data bias in a world designed for men*. Abrams, 2019.

[60] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[61] Mascha Kurpicz-Briki and Tomaso Leoni. A world full of stereotypes? further investigation on origin and gender bias in multi-lingual word embeddings. *Frontiers in big Data*, 4:20, 2021.

[62] Franziska Moser and Bettina Hannover. How gender fair are German schoolbooks in the twenty-first century? An analysis of language and illustrations in schoolbooks for mathematics and German. *European journal of psychology of education*, 29(3):387–407, 2014.

[63] Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, 2018.

[64] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65. Association for Computational Linguistics, June 2014.

[65] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, page 625–635. International World Wide Web Conferences Steering Committee, 2015.

[66] Yating Zhang, Adam Jatowt, Sourav S. Bhowmick, and Katsumi Tanaka. The past is not a foreign country: Detecting semantically similar terms across time. *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2793–2807, 2016.

[67] Robert Bamler and Stephan Mandt. Dynamic word embeddings. In *International conference on Machine learning*, pages 380–389. PMLR, 2017.

[68] Steffen Eger and Alexander Mehler. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 52–58. Association for Computational Linguistics, August 2016.

[69] Hamed Jelodar, Yongli Wang, Rita Orji, and Shucheng Huang. Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2733–2742, 2020.

[70] Ranganathan Chandrasekaran, Vikalp Mehta, Tejali Valkunde, and Evangelos Moustakas. Topics, trends, and sentiments of tweets about the covid-19 pandemic: Temporal infoveillance study. *Journal of medical Internet research*, 22(10):e22624, 2020.

[71] C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225, May 2014.

[72] Xueting Wang, Canruo Zou, Zidian Xie, and Dongmei Li. Public opinions towards covid-19 in california and new york on twitter. *MedRxiv*, 2020. https://www.medrxiv.org/content/early/2020/07/14/2020.07.12.20151936.

[73] Hyeju Jang, Emily Rempel, David Roth, Giuseppe Carenini, and Naveed Zafar Janjua. Tracking covid-19 discourse on twitter in north america: Infodemiology study using topic modeling and aspect-based sentiment analysis. *Journal of medical Internet research*, 23(2):e25431, 2021.

[74] Oren Pereg, Daniel Korat, Moshe Wasserblat, Jonathan Mamou, and Ido Dagan. Absapp: A portable weakly-supervised aspect-based sentiment extraction system. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 1–6, 2019.

[75] Danny Valdez, Marijn Ten Thij, Krishna Bathina, Lauren A Rutter, and Johan Bollen. Social media insights into us mental health during the covid-19 pandemic: Longitudinal analysis of twitter data. *Journal of medical Internet research*, 22(12):e21418, 2020.

[76] Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[77] Lin Miao, Mark Last, and Marina Litvak. Twitter data augmentation for monitoring public opinion on COVID-19 intervention measures. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Association for Computational Linguistics, December 2020.

[78] Yongtai Liu, Joshua Maynez, Gonçalo Simões, and Shashi Narayan. Data augmentation for low-resource dialogue summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 703–710. Association for Computational Linguistics, July 2022.

[79] Jia Xue, Junxiang Chen, Ran Hu, Chen Chen, Chengda Zheng, Yue Su, and Tingshao Zhu. Twitter discussions and emotions about the covid-19 pandemic: Machine learning approach. *Journal of medical Internet research*, 22(11):e20550, 2020.

[80] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[81] Klaifer Garcia and Lilian Berton. Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa. *Applied soft computing*, 101:107057, 2021.

[82] Raj Kumar Gupta and Yinping Yang. Crystalfeel at semeval-2018 task 1: Understanding and detecting emotion intensity using affective lexicons. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 256–263, 2018.

[83] Mohammad Al-Ramahi, Ahmed Elnoshokaty, Omar El-Gayar, Tareq Nasralah, and Abdullah Wahbeh. Public discourse against masks in the covid-19 era: Infodemiology study of twitter data. *JMIR Public Health and Surveillance*, 7(4):e26780, 2021.

[84] Neil Yeung, Jonathan Lai, and Jiebo Luo. Face off: Polarized public opinions on personal face mask usage during the covid-19 pandemic. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 4802–4810. IEEE, 2020.

[85] Naw Safrin Sattar and Shaikh Arifuzzaman. Covid-19 vaccination awareness and aftermath: Public sentiment analysis on twitter data and vaccinated population prediction in the usa. *Applied Sciences*, 11(13):6128, 2021.

[86] Goran Muric, Yusong Wu, and Emilio Ferrara. Covid-19 vaccine hesitancy on social media: Building a public twitter data set of antivaccine content, vaccine misinformation, and conspiracies. *JMIR public health and surveillance*, 7(11):e30642, 2021.

[87] Xupin Zhang, Hanjia Lyu, and Jiebo Luo. Understanding the hoarding behaviors during the covid-19 pandemic using large scale social media data. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 5007–5013. IEEE, 2021.

[88] Yipeng Zhang, Hanjia Lyu, Yubao Liu, Xiyang Zhang, Yu Wang, and Jiebo Luo. Monitoring depression trends on twitter during the covid-19 pandemic: Observational study. *JMIR infodemiology*, 1(1):e26769, 2021.

[89] Viet Duong, Jiebo Luo, Phu Pham, Tongyu Yang, and Yu Wang. The ivory tower lost: How college students respond differently than the general public to the covid-19 pandemic. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 126–130. IEEE, 2020.

[90] Janet M Box-Steffensmeier and Laura Moses. Meaningful messaging: Sentiment in elite social media communication with the public on the covid-19 pandemic. *Science Advances*, 7(29):eabg2898, 2021.

[91] Rashid Khan, Furqan Rustam, Khadija Kanwal, Arif Mehmood, and Gyu Sang Choi. Us based covid-19 tweets sentiment analysis using textblob and supervised machine learning algorithms. In *2021 international conference on artificial intelligence (ICAI)*, pages 1–8. IEEE, 2021.

[92] Liviu-Adrian Cotfas, Camelia Delcea, Rareș Gherai, and Ioan Roxin. Unmasking people's opinions behind mask-wearing during covid-19 pandemic—A Twitter Stance Analysis. *Symmetry*, 13(11):1995, 2021.

[93] Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. Stance detection in covid-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, volume 1, 2021.

[94] Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England journal of medicine*, 372(9):793–795, 2015.

[95] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, and Martin Landray. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[96] William B Kannel and Daniel L McGee. Diabetes and cardiovascular disease: the framingham study. *JAMA*, 241(19):2035–2038, 1979.

[97] Madeleine P Ball, Jason R Bobe, Michael F Chou, and et al. Harvard personal genome project: lessons from participatory public research. *Genome medicine*, 6(2):10, 2014.

[98] Yvette C Cozier, Julie R Palmer, and Lynn Rosenberg. Comparison of methods for collection of dna samples by mail in the black women's health study. *Annals of epidemiology*, 14(2):117–122, 2004.

[99] Isaac S Kohane. Using electronic health records to drive discovery in disease genomics. *Nature Reviews Genetics*, 12(6):417, 2011.

[100] Xiangqun Zheng-Bradley and Paul Flicek. Applications of the 1000 genomes project resources. *Briefings in functional genomics*, 16(3):163–170, 2016.

[101] George M Church. The personal genome project. *Molecular systems biology*, 1(1), 2005.

[102] Clare Turnbull, Richard H Scott, Ellen Thomas, Louise Jones, Nirupa Murugaesu, Freya Boardman Pretty, Dina Halai, Emma Baple, Clare Craig, Angela Hamblin, Shirley Henderson, Christine Patch, Amanda O'Neill, Andrew Devereau, Katherine Smith, Antonio Rueda Martin, Alona Sosinsky, Ellen M McDonagh, Razvan Sultana, Michael Mueller, et al. The 100 000 genomes project: bringing whole genome sequencing to the nhs. *BMJ*, 361, 2018.

[103] Tim C Peakman and Paul Elliott. The UK Biobank sample handling and storage validation studies. *International journal of epidemiology*, 37(suppl_1):i2–i6, 2008.

[104] Sara Chandros Hull, Richard R Sharp, and Jeffrey R Botkin. Patients' views on identifiability of samples and informed consent for genetic research. *The American Journal of Bioethics*, 8(10):62–70, 2008.

[105] Bradley Malin, David Karp, and Richard H Scheuermann. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. *Journal of Investigative Medicine*, 58(1):11–18, 2010.

[106] Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. A systematic review of re-identification attacks on health data. *PloS one*, 6(12):e28071, 2011.

[107] Latanya Sweeney, Akua Abu, and Julia Winn. *Identifying Participants in the Personal Genome Project by Name*. Data Privacy Lab, IQSS, Harvard University, April 2013.

[108] Zhiyu Wan, Yevgeniy Vorobeychik, Weiyi Xia, Ellen Wright Clayton, Murat Kantarcioglu, and Bradley Malin. Expanding access to large-scale genomic data while promoting privacy: a game theoretic approach. *The American Journal of Human Genetics*, 100(2):316–22, 2017.

[109] M.D. Choudhury and S De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, pages 71–80, 01 2014.

[110] Sairam Balani and Munmun De Choudhury. Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1373–1378, 2015.

[111] Latanya Sweeney. Simple demographics often identify people uniquely. Unpublished manuscript. Retrieved from http://dataprivacylab.org/projects/identifiability/, 2000.

[112] Wikipedia contributors. Category: Cohort studies. *Wikipedia*, 2016. https://en.wikipedia.org/wiki/Category:Cohort_studies.

[113] Muhammad Naveed, Erman Ayday, Ellen W Clayton, Jacques Fellay, Carl A Gunter, Jean-Pierre Hubaux, Bradley A Malin, and XiaoFeng Wang. Privacy in the genomic era. *ACM Computing Surveys (CSUR)*, 48(1):1–44, 2015.

[114] Grigorios Loukides, Joshua C Denny, and Bradley Malin. The disclosure of diagnosis codes can breach research participants' privacy. *Journal of the American Medical Informatics Association*, 17(3):322–327, 2010.

[115] Zhen Lin, Art B. Owen, and Russ B. Altman. Genomic research and human subject privacy. *Science*, 305(5681):183–183, 2004.

[116] Khaled El Emam, Sam Rodgers, and Bradley Malin. Anonymising and sharing individual patient data. *BMJ (Clinical research ed.)*, 350:h1139, 2015.

[117] Bradley Malin and Latanya Sweeney. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of biomedical informatics*, 37(3):179–192, 2004.

[118] Ravi V Atreya, Joshua C Smith, Allison B McCoy, Bradley Malin, and Randolph A Miller. Reducing patient re-identification risk for laboratory results within research datasets. *Journal of the American Medical Informatics Association*, 20(1):95–101, 2013.

[119] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008.

[120] Melissa Gymrek, Amy L McGuire, David Golan, Eran Halperin, and Yaniv Erlich. Identifying personal genomes by surname inference. *Science*, 339(6117):321–324, 2013.

[121] Sriram Sankararaman, Guillaume Obozinski, Michael I Jordan, and Eran Halperin. Genomic privacy and limits of individual detection in a pool. *Nature genetics*, 41(9):965–967, 2009.

[122] Suyash S Shringarpure and Carlos D Bustamante. Privacy risks from genomic data-sharing beacons. *The American Journal of Human Genetics*, 97(5):631–646, 2015.

[123] Jean Louis Raisaro, Florian Tramer, Zhanglong Ji, Diyue Bu, Yongan Zhao, Knox Carey, David Lloyd, Heidi Sofia, Dixie Baker, Paul Flicek, et al. Addressing beacon re-identification attacks: quantification and mitigation of privacy risks. *Journal of the American Medical Informatics Association*, 24(4):799–805, 2017.

[124] Zhiyu Wan, Yevgeniy Vorobeychik, Murat Kantarcioglu, and Bradley Malin. Controlling the signal: Practical privacy protection of genomic data sharing through beacon services. *BMC medical genomics*, 10(2):87–100, 2017.

[125] Joshua C Denny, Marylyn D Ritchie, Melissa A Basford, Jill M Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R Masys, Dan M Roden, and Dana C Crawford. Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9):1205–1210, 2010.

[126] Alexandros Mittos, Jeremy Blackburn, and Emiliano De Cristofaro. "23andme confirms: I'm super white"–analyzing twitter discourse on genetic testing. *ArXiv*, 2018. https://arxiv.org/abs/1801.09946.

[127] Monika Taddicken. The "privacy paradox" in the social web: The impact of privacy concerns, individual characteristics, and the perceived social relevance on different forms of self-disclosure. *Journal of Computer-Mediated Communication*, 19(2):248–273, 2014.

[128] Elizabeth M Morgan, Chareen Snelson, and Patt Elison-Bowers. Image and video disclosure of substance use on social media websites. *Computers in Human Behavior*, 26(6):1405–1411, 2010.

[129] Carl L Hanson, Scott H Burton, Christophe Giraud-Carrier, Josh H West, Michael D Barnes, and Bret Hansen. Tweaking and tweeting: exploring twitter for nonmedical use of a psychostimulant drug (adderall) among college students. *Journal of Medical Internet Research*, 15(4):e62, 2013.

[130] Carl Lee Hanson, Ben Cannon, Scott Burton, and Christophe Giraud-Carrier. An exploration of social circles and prescription drug abuse through twitter. *Journal of medical Internet research*, 15(9):e189, 2013.

[131] Huina Mao, Xin Shuai, and Apu Kapadia. Loose tweets: an analysis of privacy leaks on twitter. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, pages 1–12, 2011.

[132] Emily Christofides, Amy Muise, and Serge Desmarais. Information disclosure and control on facebook: Are they two sides of the same coin or two different processes? *Cyberpsychology & behavior*, 12(3):341–345, 2009.

[133] Laura J Rasmussen-Torvik, Sarah C Stallings, Adam S Gordon, Berta Almoguera, Melissa A Basford, Suzette J Bielinski, Ariel Brautbar, MH Brilliant, David S Carrell, JJ Connolly, et al. Design and anticipated outcomes of the emerge-pgx project: a multicenter pilot for preemptive pharmacogenomics in electronic health record systems. *Clinical Pharmacology & Therapeutics*, 96(4):482–489, 2014.

[134] Dan M Roden, Jill M Pulley, Melissa A Basford, Gordon R Bernard, Ellen W Clayton, Jeffrey R Balser, and Dan R Masys. Development of a large-scale de-identified dna biobank to enable personalized medicine. *Clinical Pharmacology & Therapeutics*, 84(3):362–369, 2008.

[135] Raymond Heatherly, Joshua C Denny, Jonathan L Haines, Dan M Roden, and Bradley A Malin. Size matters: How population size influences genotype–phenotype association studies in anonymized data. *Journal of biomedical informatics*, 52:243–250, 2014.

[136] Jane Elliott and Peter Shepherd. Cohort profile: 1970 british birth cohort (bcs70). *International journal of epidemiology*, 35(4):836–843, 2006.

[137] Stephen S Rich, Patrick Concannon, and Henry Erlich. The type 1 diabetes genetics consortium. *Annals of the New York Academy of Sciences*, 1079(1):1–8, 2006.

[138] Edward Giovannucci, Meir J Stampfer, Graham A Colditz, et al. Multivitamin use, folate, and colon cancer in women in the nurses' health study. *Annals of internal medicine*, 129(7):517–524, 1998.

[139] V Beral and Million Women Study Collaborators. Breast cancer and hormone-replacement therapy in the million women study. *The Lancet*, 362(9382):419–427, 2003.

[140] Emelia J Benjamin, Philip A Wolf, Ralph B D'Agostino, Halit Silbershatz, William B Kannel, and Daniel Levy. Impact of atrial fibrillation on the risk of death: the framingham heart study. *Circulation*, 98(10):946–952, 1998.

[141] Hanan Al Kuwari, Asma Al Thani, and Ajayeb Al Marri. The qatar biobank: background and methods. *BMC public health*, 15(1):1208, 2015.

[142] International HapMap Consortium. The international hapmap project. *Nature*, 426(6968):789–96, 2003.

[143] Zhijun Yin, You Chen, Daniel Fabbri, Jimeng Sun, and Bradley Malin. #prayfordad: learning the semantics behind why social media users disclose health information. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, page 456–465. International AAAI Conference on Weblogs and Social Media, 2016.

[144] A Machanavajjhala, J Gehrke, D Kifer, and M Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*, page 24. IEEE, 2006.

[145] Peter M Visscher and William G Hill. The limits of individual identification from sample allele frequencies: theory and statistical analysis. *PLoS genetics*, 5(10):e1000628, 2009.

[146] Kris A. Wetterstrand. The cost of sequencing a human genome. *Genome.gov*, November 2021. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.

[147] Cecelia A Bellcross, Patricia Z Page, and Dana Meaney-Delman. Direct-to-consumer personal genome testing and cancer risk prediction. *The Cancer Journal*, 18(4):293–302, 2012.

[148] Antonio Regalado. More than 26 million people have taken an at-home ancestry test. *MIT Technology Review*, June 2020. https://www.technologyreview.com/2019/02/11/103446/more-than-26-million-people-have-taken-an-at-home-ancestry-test/.

[149] Marc McDermott. 23andme vs ancestrydna. *SmarterHobby*, January 2022. https://www.smarterhobby.com/genealogy/23andme-vs-ancestry-dna/.

[150] Tobias Haeusermann, Bastian Greshake, Alessandro Blasimme, Darja Irdam, Martin Richards, and Effy Vayena. Open sharing of genomic data: Who does it and why? *PLoS One*, 12(5):e0177158, 2017.

[151] Haitao Xu, Haining Wang, and Angelos Stavrou. Privacy risk assessment on online photos. In *International Symposium on Recent Advances in Intrusion Detection*, pages 427–447. Springer, 2015.

[152] Rajagopal Venkatesaramani, Bradley A Malin, and Yevgeniy Vorobeychik. Re-identification of individuals in genomic datasets using public face images. *Science advances*, 7(47):eabg3296, 2021.

[153] Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. Deep face recognition: A survey. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pages 471–478. IEEE, 2018.

[154] Martin Bauml, Makarand Tapaswi, and Rainer Stiefelhagen. Semi-supervised learning with constraints for person identification in multimedia data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3602–3609, 2013.

[155] Shareen Irshad and Tariq Rahim Soomro. Identity theft and social media. *International Journal of Computer Science and Network Security*, 18(1):43–55, 2018.

[156] Alessandro Acquisti and Christina Fong. An experiment in hiring discrimination via online social networks. *Management Science*, 66(3):1005–1024, 2020.

[157] Amanda Nosko, Eileen Wood, and Seija Molema. All about me: Disclosure in online social networking profiles: The case of facebook. *Computers in human behavior*, 26(3):406–418, 2010.

[158] Zhiyu Wan, Yevgeniy Vorobeychik, Weiyi Xia, Ellen Wright Clayton, Murat Kantarcioglu, Ranjit Ganta, Raymond Heatherly, and Bradley A Malin. A game theoretic framework for analyzing re-identification risk. *PloS one*, 10(3):e0120592, 2015.

[159] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.

[160] Patricia Maranga. Social photos generate more engagement: New research. *Social Media Marketing — Social Media Examiner*, May 2014. https://www.socialmediaexaminer.com/photos-generate-engagement-research/.

[161] Saeideh Bakhshi, David A Shamma, and Eric Gilbert. Faces engage us: Photos with faces attract more likes and comments on instagram. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 965–974, 2014.

[162] Mahmoud Elbattah., Émilien Arnaud., Maxime Gignon., and Gilles Dequen. The role of text analytics in healthcare: A review of recent developments and applications. In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies - Scale-IT-up,*, pages 825–832. INSTICC, SciTePress, 2021.

[163] J Scott Roberts, Michele C Gornick, Deanna Alexis Carere, Wendy R Uhlmann, Mack T Ruffin, and Robert C Green. Direct-to-consumer genetic testing: user motivations, decision making, and perceived utility of results. *Public Health Genomics*, 20(1):36–45, 2017.

[164] European Society of Human Genetics. Statement of the eshg on direct-to-consumer genetic testing for health-related purposes. *European Journal of Human Genetics*, 18(12):1271, 2010.

[165] Ellen W Clayton, Colin M Halverson, Nila A Sathe, and Bradley A Malin. A systematic literature review of individuals' perspectives on privacy and genetic information in the united states. *PLoS One*, 13(10):e0204417, 2018.

[166] Lukasz Olejnik, Agnieszka Kutrowska, and Claude Castelluccia. The beginning of genetic exhibitionism? In *Proceedings of the 1st Workshop on Genome Privacy*, 2014.

[167] Zhijun Yin, Lijun Song, and Bradley Malin. Reciprocity and its association with treatment adherence in an online breast cancer forum. In *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 618–623. IEEE, 2017.

[168] Wenjing Pan, Bo Feng, Cuihua Shen, et al. Examining social capital, social support, and language use in an online depression forum: Social network and content analysis. *Journal of medical Internet research*, 22(6):e17365, 2020.

[169] John A Naslund, Ameya Bondre, John Torous, and Kelly A Aschbrenner. Social media and mental health: benefits, risks, and opportunities for research and practice. *Journal of technology in behavioral science*, 5(3):245–257, 2020.

[170] Alessandro Acquisti. Privacy in electronic commerce and the economics of immediate gratification. In *Proceedings of the 5th ACM conference on Electronic commerce*, pages 21–29, 2004.

[171] ageitgey. Github - ageitgey/face_recognition: The world's simplest facial recognition api for python and the command line. *GitHub*, June 2022. https://github.com/ageitgey/face_recognition.

[172] John W Mohr and Petko Bogdanov. Introduction—topic models: What they are and why they matter. *Poetics*, 41(6):545–569, 2013.

[173] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[174] Jay M Ver Hoef and Peter L Boveng. Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–2772, 2007.

[175] Patricia Redsicker. Social photos generate more engagement: New research. *Social media examiner*, 50(1), 2014.

[176] Nathan C Lindstedt. Structural topic modeling for social scientists: a brief case study with social movement studies literature, 2005–2017. *Social Currents*, 6(4):307–318, 2019.

[177] Rajagopal Venkatesaramani, Doug Downey, Bradley Malin, and Yevgeniy Vorobeychik. A semantic cover approach for topic modeling. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*, pages 92–102, 2019.

[178] Yongcheng Zhan, Ruoran Liu, Qiudan Li, Scott James Leischow, and Daniel Dajun Zeng. Identifying topics for e-cigarette user-generated contents: a case study from multiple social media platforms. *Journal of medical Internet research*, 19(1):e5780, 2017.

[179] CDC. COVID Data Tracker. *Centers for Disease Control and Prevention*, March 2020. https://covid.cdc.gov/covid-data-tracker.

[180] Weber Lauren. Covid is killing rural americans at twice the rate of people in urban areas. *NBC News*, June 2022.

[181] Diego F. Cuadros, Adam J. Branscum, Zindoga Mukandavire, F. DeWolfe Miller, and Neil MacKinnon. Dynamics of the COVID-19 epidemic in urban and rural areas in the United States. *Annals of Epidemiology*, 59:16–20, July 2021.

[182] Rishabh Singh Chauhan, Denise Capasso da Silva, Deborah Salon, Ali Shamshiripour, Ehsan Rahimi, Uttara Sutradhar, Sara Khoeini, Abolfazl (Kouros) Mohammadian, Sybil Derrible, and Ram Pendyala. COVID-19 related Attitudes and Risk Perceptions across Urban, Rural, and Suburban Areas in the United States. *Findings*, page 23714, June 2021.

[183] Donald J. Alcendor. Targeting COVID Vaccine Hesitancy in Rural Communities in Tennessee: Implications for Extending the COVID-19 Pandemic in the South. *Vaccines*, 9(11):1279, November 2021.

[184] Jon Green, Jared Edgerton, Daniel Naftel, Kelsey Shoub, and Skyler J Cranmer. Elusive consensus: Polarization in elite communication on the covid-19 pandemic. *Science advances*, 6(28):eabc2717, 2020.

[185] Shana Kushner Gadarian, Sara Wallace Goodman, and Thomas B Pepinsky. Partisanship, health behavior, and policy attitudes in the early stages of the covid-19 pandemic. *Plos one*, 16(4):e0249596, 2021.

[186] Zhijun Yin, Yongtai Liu, Allison B McCoy, Bradley A Malin, Patricia R Sengstack, et al. Contribution of free-text comments to the burden of documentation: Assessment and analysis of vital sign comments in flowsheets. *Journal of medical Internet research*, 23(3):e22806, 2021.

[187] Joanne Chen Lyu, Eileen Le Han, and Garving K Luli. Covid-19 vaccine–related discussion on twitter: topic modeling and sentiment analysis. *Journal of medical Internet research*, 23(6):e24435, 2021.

[188] Yong Chen, Hui Zhang, Rui Liu, Zhiwen Ye, and Jianying Lin. Experimental explorations on short text topic mining between lda and nmf based schemes. *Knowledge-Based Systems*, 163:1–13, 2019.

[189] Alaa Abd-Alrazaq, Dari Alhuwail, Mowafa Househ, Mounir Hamdi, Zubair Shah, et al. Top concerns of tweeters during the covid-19 pandemic: infoveillance study. *Journal of medical Internet research*, 22(4):e19016, 2020.

[190] May Oo Lwin, Jiahui Lu, Anita Sheldenkar, Peter Johannes Schulz, Wonsun Shin, Raj Gupta, and Yinping Yang. Global sentiments surrounding the covid-19 pandemic on twitter: analysis of twitter trends. *JMIR public health and surveillance*, 6(2):e19447, 2020.

[191] Liviu-Adrian Cotfas, Camelia Delcea, Ioan Roxin, Corina Ioanăş, Dana Simona Gherai, and Federico Tajariol. The longest month: Analyzing covid-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. *IEEE Access*, 9:33203–33223, 2021.

[192] Emily Chen, Kristina Lerman, and Emilio Ferrara. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR public health and surveillance*, 6(2):e19273, 2020.

[193] John Cromartie. Rural-urban commuting area codes. *USDA*, August 2020. https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes.aspx.

[194] Tracy Onega, Julie E Weiss, Jennifer Alford-Teaster, Martha Goodrich, M Scottie Eliassen, and Sunny Jung Kim. Concordance of rural-urban self-identity and zip code-derived rural-urban commuting area (ruca) designation. *The Journal of Rural Health*, 36(2):274–280, 2020.

[195] Asnake Hailu. Guidelines for using rural-urban classification systems for community health assessment. *Washington State Department of Health*, October 2016. https://doh.wa.gov/sites/default/files/legacy/Documents/1500//RUCAGuide.pdf.

[196] Anqi Cui, Min Zhang, Yiqun Liu, Shaoping Ma, and Kuo Zhang. Discover breaking events with popular hashtags in twitter. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1794–1798, 2012.

[197] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

[198] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer, 2001.

[199] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.

[200] Davoud Moulavi, Pablo A Jaskowiak, Ricardo JGB Campello, Arthur Zimek, and Jörg Sander. Density-based clustering validation. In *Proceedings of the 2014 SIAM international conference on data mining*, pages 839–847. SIAM, 2014.

[201] Ibrahim Kaibi, Hassan Satori, et al. A comparative evaluation of word embeddings techniques for twitter sentiment analysis. In *2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*, pages 1–4. IEEE, 2019.

[202] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), May 2010.

[203] Xiao Yang, Craig Macdonald, and Iadh Ounis. Using word embeddings in twitter election classification. *Information Retrieval Journal*, 21(2):183–207, 2018.

[204] National Center for Health Statistics. Data Access - Urban Rural Classification Scheme for Counties. *Centers for Disease Control and Prevention*, December 2019. https://www.cdc.gov/nchs/data_access/urban_rural.htm.

[205] Ben Harris and Neil Mehrotra. The Data Underlying America's Strong Economic Recovery. *U.S. Department of the Treasury*, June 2022. https://home.treasury.gov/news/featured-stories/the-data-underlying-americas-strong-economic-recovery.

[206] Yongtai Liu. COVID19-W2V. *GitHub*, August 2022. https://github.com/yongtai123/COVID19-W2V.

[207] James G Gimpel, Nathan Lovin, Bryant Moy, and Andrew Reeves. The urban–rural gulf in american political behavior. *Political Behavior*, 42(4):1343–1368, 2020.

[208] Austin C Kozlowski, Matt Taddy, and James A Evans. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949, 2019.

[209] Mingxuan Ju, Wei Song, Shiyu Sun, Yanfang Ye, Yujie Fan, Shifu Hou, Kenneth Loparo, and Liang Zhao. Dr. emotion: Disentangled representation learning for emotion analysis on social media to improve community resilience in the covid-19 era and beyond. In *Proceedings of the Web Conference 2021*, pages 518–528, 2021.

[210] Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, 5(1):30–43, 2018.

[211] Wei Fang, Moin Nadeem, Mitra Mohtarami, and James Glass. Neural multi-task learning for stance prediction. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 13–19, 2019.

[212] Jing Ma, Wei Gao, and Kam-Fai Wong. Detect rumor and stance jointly by neural multi-task learning. In *Companion proceedings of the the web conference 2018*, pages 585–593, 2018.