Characterizing Misinformed Online Health Communities

Shahan Ali Memon

6th August 2020

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Kathleen M. Carley (Chair), Carnegie Mellon University
Bhiksha Raj, Carnegie Mellon University
David A. Broniatowski, George Washington University

This thesis document is submitted in partial fulfillment of the requirements for the degree of Master of Science.

This research is funded by Carnegie Mellon University's Center of Machine Learning for Health (CMLH) under the CMLH Fellowships in Digital Health.

Keywords: misinformation, disinformation, anti-vaccination, network science, network analysis, complex networks, network, network community structure, sociolinguistics

Iqra bismi rab bikal lazee khalaq
Khalaqal insaana min 'alaq
Iqra wa rab bukal akram
Al lazee 'allama bil qalam
'Al lamal insaana ma lam y'alam
Kallaa innal insaana layatghaa
Ar-ra aahus taghnaa
Innna ilaa rabbikar ruj'aa
Ara-aital lazee yanhaa
'Abdan iza sallaa
Ara-aita in kana 'alal hudaa
Au amara bit taqwaa
Ara-aita in kaz zaba wa ta walla
Alam y'alam bi-an nal lahaa yaraa
Kalla la illam yantahi la nasfa'am bin nasiyah
Nasiyatin kazi batin khaatiyah
Fal yad'u naadiyah
Sanad 'uz zabaaniyah
Kalla; la tuti'hu wasjud waqtarib

To my family

Abstract

Do vaccines cause Autism? Does drinking bleach cure coronavirus? Is polio vaccination a ploy to sterilize and reduce the population? From a medical perspective, the definitive answer to all these questions would be "no". And yet, the web is populated by a cacophony of mixed opinions about these issues triggered by the proliferation of public health misinformation.

Health-related misinformation has detrimental effects on the public health, and debunking it is a challenging task. Because these misinformed sub-communities discourage differing beliefs, public health practitioners and policy makers must grapple with the challenge of *penetrating* into these communities to disseminate facts or conduct any message-based intervention.

Combating the spread of false information by *differential promotion* or *censorship* of the content, or by *broadcasting* facts does not work. Instead, there is a need to strategically communicate with the misinformed communities. This requires a thorough understanding of what is an effective communication paradigm to debunk such myths.

For an effective message-based intervention, it is imperative to focus on *preference-based framing* where the preferences of the target sub-community are taken into consideration. These preferences can be defined over two main aspects: (i) *who* should deliver the message; (ii) *what* should the message be. Choosing the right messenger(s) requires understanding of how these online communities interact by tapping into their network structures. Choosing the content of the message, on the other hand, requires a thorough understanding of what *language choices* they make, and how those language choices reflect their *non-negotiable social identities*.

In this work, we identify two different health communities online: (i) vaccination sub-communities; and (ii) COVID-19 misinformation sub-communities. In the first part of this thesis, we characterize the network, and sociolinguistic variation in the online competing vaccination sub-communities to understand their linguistic choices and motivations. With the emergence of COVID-19 pandemic, the political and medical misinformation has elevated to create what is being commonly referred to as the *global infodemic*. Thus, in the second part, we first introduce a novel Twitter dataset, *CMU-MisCov19* annotated for different COVID-19 themes. We then use this dataset to characterize the competing COVID-19 misinformation sub-communities.

Our analyses show that the competing sub-communities within each part tend to have significant differences in their communication patterns, and that these differences can be leveraged to form better message interventions. We also make our annotated dataset available for the community to use for further analysis.

Acknowledgments

I would like to first thank Carnegie Mellon University's Center for Machine Learning and Health (CMLH) for their fellowship to support this work.

I would then especially thank my advisor, Prof. Kathleen M. Carley, for sharing with me her vast knowledge on social network analysis, and for her hours of advice and mentorship via our weekly meetings.

I would like to thank my advisors and mentors, Prof. Rita Singh, Prof. Bhiksha Raj, and Dr. Ingmar Weber for their continuous support throughout my time at Carnegie Mellon.

I thank David A. Broniatowski for accepting my invitation to be on my thesis committee, and for his advice and suggestions.

I am also grateful to David R. Mortensen, Lori Levin, Robert E. Frederking, Kate Schaich, and the members of Computational Analysis of Social and Organizational Systems (CASOS), especially Aman Tyagi, David Beskow, and Matthew Babcock for their advice, support, and help with different tools and resources throughout this project.

I would also like to recognize my friends and roommates, Sannan Tariq, Muhammad Ahmed Shah, Hira Dhamyal, Vanessa Fernandes, and Aqsa Kashaf for all our amazing discussions, and their support throughout my time here.

Last, but not the least, I dedicate all my past and future work to my family: Arfana Shaikh (my mother), Imtiaz Ahmed (my father), Assad Ali (my brother), Rumasa Pareesha (my sister), Riyan Mujtaba (my brother), Azka Qaiser (my sister-in-law), Mustafa Memon (my nephew), Zaibunissa Shaikh (my grandmother), Nizam Shaikh (my uncle), and Afroz Shaikh (my Aunt); thank you for your selfless efforts to support my studies and research, and for always believing in me.

Contents

Chapter 1

Introduction

Public health misinformation can be defined as a *health-related claim that is currently false due to a lack of scientific evidence* [34]. Health-related misinformation has detrimental effects on the public health. According to researchers, many preventable diseases have re-emerged as a consequence of the drop in immunization rates due to declining trust in vaccines caused by the misinformation on the web [62]. According to the Center of Disease Control and Prevention (CDC), measles - which was declared to have been eliminated from the United States in 2000 [4] - re-emerged in places such as Portland [52], Boston [74], Chicago [2], and Michigan [52]. In fact, a staggering 30% increase was seen in measles cases [3] between 2016 and 2017 most of which is attributed to online health-related myths. A more recent example of the life-threatening effects of misinformation would be what many are referring to as the COVID-19 *infodemic* [35]. The emergence of COVID-19 pandemic has also given rise to conspiracy theories, false cures, false preventions, and false treatments. Because the quality of information people receive affects their perception which in turn affects their actions [14], COVID-19 misinformation is bound to undermine the efforts to limit the virus [12]. Another recent research found that Twitter bots were sharing content that contributed to positive sentiments about e-cigarettes in the U.S [62]. The U.S. is not the only country to face such an issue. Water fluoridation myths increasing incidence of tooth decay in children in Australia [13], surge of measles cases due to anti-vaccination campaigns in Italy [75], and increase in Ebola death toll in West Africa due to conspiracy theories [7] are only a few examples of the negative influence of health misinformation in other countries.

This makes debunking of false information vitally important. According to one study [85], if left undisputed, misinformation can in fact exacerbate the spread of the epidemic itself. Process of debunking misinformation, however, is complex and one that is not completely understood [32]. This is because in order to conduct any intervention, it is first imperative to be able to identify the misinformation, as well as the misinformed communities. Because of the scarcity of data, and diversity of misinformation themes, this is already a challenging task in itself, but is also not enough. A second, and arguably a more important aspect of an intervention is to be able to *correct* and *change* the beliefs of the misinformed communities. To be able to do this, it is important to understand how different communities interact, and what is the right and the most effective *communication paradigm* to conduct an intervention.

Two of the main aspects of a communication paradigm for a message-based intervention are: (i) *who* is/are the right messenger(s), and (ii) *what* is the right message.

Many online communities, such as anti-vaxxers, thrive not on evidence-based proofs but rather on *social proofs* where it matters a lot who the messenger is and how popular the message is in the local circle [1]. It may be much more effective, therefore, if the message is received from a known (*and trusted*) person rather than a purported untrustworthy government authority [90]. It is therefore important to identify the right messenger(s). This requires understanding how different communities are connected to each other, and what patterns of communication are the most effective.

But message-based interventions fall flat if the *framing* of the message and the *message* itself are not persuasive enough. For an effective health communication, it is imperative to focus on *preference-based framing* where the preferences of the target sub-community are taken into consideration to create an effective content for the message. This requires a thorough understanding of what *language choices* these communities make, and how those language choices reflect their *non-negotiable social identities*.

The research question thus is *"What is an effective message, framing, and the messenger to debunk public health misinformation in the online communities"*

In this work we conduct observational studies on understanding different health communities online using two case studies. We characterize these communities in terms of their communication network and linguistic patterns. Based on our analyses, we make some *suggestions* for the public health practitioners to follow.

We focus on two online Twitter communities: In the first part of this thesis, we characterize vaccination communities. This is because vaccination-related misinformation happens to be one of the most prevalent and long-standing form of misinformation. Due to the recent spread of COVID-19, false information has hampered proper communication leading to an *infodemic*. Therefore, in the second part of this thesis, we first collect and annotate a novel Twitter COVID-19 dataset, called *CMU-MisCOV19*, and then use it to characterize COVID-19 misinformation communities.

In all, there are two questions that this thesis is trying to address: (i) are there social and linguistic differences in those that take a different stance on a topic?; and (ii) are there different linguistic features to different types of misinformation or accurate information? Many issues such as climate change or vaccination are directed enough to trigger competing beliefs (i.e. pro- and anti), and so many individuals, though not all, take stances based on what resonates with their beliefs. The relationship between disinformation and stance vis-a-vis issues, however, is complex, and the stance around a particular topic isn't always clearly defined. An example would be COVID-19 discourse which is an amalgamation of many sub-topics each of which can have a stance. Because in this case, the stance of a particular agent is not clear, study of the discourse around COVID-19 warrants a different strategy. Consequently, in this thesis, we define and study communities and the discourse around misinformed communities in two different ways: In the first part, we define communities based on stance, and in the second part, we define communities based on the propagation and endorsement of misinformation. Eventually, our hope is to develop a systematic way of characterizing communities both in terms of stance and misinformation. This thesis only lays the ground work for this, by exploring the linguistic and social network features relative to these two ways of addressing the conversation around an issue.

This thesis document is organized as follows: Chapter 2 describes in detail our analysis on the sociolinguistic and network variation between the two competing vaccination communities;

Chapter 3 describes our COVID-19 misinformation data collection strategy, and presents preliminary analysis on characterizing COVID-19 misinformation communities using our novel Twitter dataset; and finally, chapter 4 describes some takeaways, and future work pertaining to our work.

Part I

Exploring Vaccination Communities

Chapter 2

Characterizing Sociolinguistic Variation in the Competing Vaccination Communities

Vaccination related misinformation is arguably the most prevalent form of misinformation online. Therefore, for the purposes of this study, we chose to tap into vaccination discourse on Twitter. We study the conversations between two competing groups of Twitter users: (i) those who believe in the effectiveness of vaccinations *(pro-vaxxers)*, and (ii) those who are skeptical *(anti-vaxxers)*. The goal of our study was to characterize the two competing vaccination communities in terms of their sociolinguistic, and network variation. Our data analysis show significant linguistic variation between the two communities in terms of their usage of linguistic intensifiers, pronouns, and uncertainty words. Our network-level analysis show significant differences between the two communities in terms of their network density, echo-chamberness, and the EI index. We hypothesize that these sociolinguistic differences can be used as proxies to characterize and understand these communities to devise better message interventions.[1]

2.1 Background Literature

A large chunk of related work falls in the category of understanding the attitudes of anti-vaccination (and vaccination) advocates, or more concretely in uncovering *why* people have certain beliefs. This line of work is somewhat spread around opinion pieces, and topic modeling strategies to understand themes around vaccine communication in online sub-communities. A general methodology for all these studies [41, 49, 53, 58, 67, 95, 97] is to first identify the different sub-communities, and then compare them within a set of variables (e.g. network structure, communication, topics, themes, etc.).

Different anti-vaccination themes have become drivers of vaccine refusal. These have been identified in the literature with varying methodologies. These themes include safety and effectiveness [41, 53, 97]; alternate medicine [53, 58]; lack of trust [41, 53, 97]; civil liberties [58]; conspiracy theories [53, 58]; morality, religion and ideology [58, 96]; misinformation and falsehood [41, 97]; emotive appeals [58]; advocacy for natural healing [41, 58]; and content aspects [58].

[1]This work was accepted at SBP-BRiMS 2020, a preprint of which can be found at [66]

As observed, misinformation is just *one* of the many factors that influence the online vaccine communication. Nevertheless, it is one of the most widely studied themes in terms of anti-vaccination, partly because most of the recent measles cases have been directly attributed to misinformation. Also misinformation is a broad category that is triggered by conspiracy theories [57, 58], myths [63], false scientific evidences [58], misconceptions [40], and attribution errors in narratives of other misinformed users [80].

Themes and topics are not the only ways to characterize anti-vaxxers. Researchers in [50, 83] argue that anti-vaccination beliefs can be projected over a wide spectrum with different levels of valence: vaccine rejectors (VRj) are anti-vaxxers entrenched in their refusal due to misinformation; vaccine resistant (VR) are those that are still open to listening; and vaccine hesitant (VH) are those who tend to have anxiety about vaccines, but are not committed to vaccine refusal.

In our work [66], however, we simplify our clustering process by assigning individuals in two competing groups: *pro-vaxxers*, and *anti-vaxxers*. We then look at the network and sociolinguistic variation around these two competing vaccination communities to understand their communication patterns. In this regard, prior work includes the sociolinguistic analysis of Twitter in multilingual societies [59], predicting community membership using word frequencies [28], and identifying effective vaccine communication using fuzzy trace theory [24]. The most relevant study to our work is by Duseja and Jhamtani in [43] on the sociolinguistic study of online echo-chambers [43]. We specifically apply their work to vaccination communities to understand their differences in usage of linguistic intensifiers, pronouns, and uncertainty words. We also conduct a network-level analysis by computing the network density, EI index, and echo-chamberness for the two target communities.

2.2 Dataset

To construct our dataset, we employ a three-stage process: (i) we first collect data using a set of hashtags via the Twitter search and the Twitter streaming API; (ii) we use this data to identify the two communities; and (iii) finally, to mitigate survivorship bias [27] and collect more data per individual, we collect timelines of the identified pro- and anti-vaxxers. We describe this process in detail in the following subsections. In the section 2.2.4, we present the statistics for the final set of data we use to conduct our analyses.

2.2.1 Data Collection

We first collect a set of known pro-vaccination and anti-vaccination hashtags from our domain knowledge as well as from the background literature [42]. List of these hashtags can be found in Table 2.1. We use these hashtags to collect Twitter data through the Twitter Streaming API, and augment it with data collected from Twitter Search API. The data consists of Tweets from 29th October 2019 to 12th November 2019. Based on [25], we filter out all tweets that do not include the lemmas "vacc" or "vax" (case insensitive) as part of their tweet text. This is to remove any possible noise in the data.

It is important to acknowledge that our data does not represent the full discourse on vaccination. This is because (i) we collect data only for a two week time-period; and (ii) most of our

hashtags are chosen to specifically extract users that are already part of pro- and anti-vaccination groups, and therefore may not capture the "undecided" users. This is a design choice as our study *does not* concern the characterization of the overall discourse, but instead only deals with the illumination of rhetorical and social differences between the two online competing groups.

Table 2.1: This table shows the hashtags used for the task of data collection. We use camel-casing for better readability.

Stance	Hashtags
Pro-vaccination	*VaccinesSaveLives, VaccinesWork, WorldImmunizationWeek, VaxWithMe, HealthForAll, WiW, ThankYouLaura*
Anti-vaccination	*LearnTheRisk, VaccineInjury, VaccineDeath, VaccineDamage, VaccinesCauseAutism, CDCFraud, CDCWhistleBlower, CDCTruth, WakeUpAmerica, HearUs, HealthFreedom*
Unidentified	*Vaccine, Vaccines, Vaccinate, VaccinateUS*

2.2.2 Community Detection

Label Propagation

To be able to conduct any analysis, it is imperative to identify the competing groups. Assigning a stance to a tweet or a twitter user is a non-trivial problem. Therefore, we use a similar method as described in [86, 87] to find anti-vaxxers and pro-vaxxer groups based on the weighted combination of the *valence* of their hashtags. In our study, we assume that retweets indicate endorsement.

In the previous studies such as [44], hashtags have been shown to work as realistic proxies for identifying stances among different groups on social media sites. In [86], hashtags are used to identify twitter users who believe in anthropogenic causes of climate change and those who do not. Similarly, in [87], hashtags could also be used to identify polarization in political discourse and how the polarization can change with time.

We use community detection method based on the work done in [87]. We first choose 2 seed hashtags for each of the polarized groups: *#VaccinesSaveLives* and *#VaccinesWork* for pro-vaccination and *#VaccineInjury* and *#LearnTheRisk* for anti-vaccination. We assign pro-vaccination seeds a valence of +1, and anti-vaccination seeds a valence of -1. [2] We then create a hashtag co-occurrence graph to identify most co-occurring hashtags with the chosen seeds, and choose those that are semantically similar, as well as the ones that are known to be pro-vax and anti-vax hashtags from the background literature [24, 25, 42] to manually assign a hard valence of +1 and -1. We then use a variant of label propagation algorithm [92] described as Algorithm 1 below to assign valence to each of the remaining hashtags. Similar to [87] the input to the algorithm is a hashtag-to-hashtag co-occurrence graph where hashtags represent nodes, and nodes are connected if they co-occur. The edges are weighted by the frequency of co-occurrence.

[2]To validate our choice of hashtags, we randomly sample 100 tweets for each of these hashtags. For pro-vaccination hashtags, 98% of tweets with hashtag *#VaccinesSaveLives* and 97% of tweets with hashtag *#VaccinesWork* were related to pro-vaccination. For anti-vaccination hashtags, 88% of tweets with hashtag *#LearnTheRisk* and 93% of tweets with hashtag *#VaccineInjury* were related to anti-vaccination.

Algorithm 1: Label Propagation Algorithm

Input: Nodes = n; Edges = e; Edge Weight = e_{ij}, $i \in n$ and $j \in n$
initialize $\gamma = 50$ and i;
for *each n* **do**
  define $l = integer(i/\gamma)$; $i+ = 1$;
  for *each n* **do**
    if *n not labeled* **then**
      compute t = neighbors of n;
      compute t_l = labeled neighbors of n;
      if $|t_l| + l \geq t$ **then**
        initialize *score*, c
        for *each* $t_i \in t$ **do**
          score += label t_i * e_{nt_i}
          c += e_{nt_i}
        end
        update label $n = score/c$
      end
    end
  end
end

Stance Identification

Once we have identified the valence of a set of hashtags, we aggregate hashtags used by each user and find a weighted average of the valence of all hashtags used by a particular user. We label a user as pro-vaxxer, or anti-vaxxer if the weighted average was positive, or negative respectively.

Using the algorithm, 3295 users are identified as pro-vaxxers, 2967 as anti-vaxxers. We randomly sample 100 users that were classified as pro-vaxxers and 100 users that were classified as anti-vaxxers to evaluate the quality of assignment. We find 96% of the labeled pro-vaxxers as pro-vaxxers, and 80% of the labeled anti-vaxxers as anti-vaxxers.

2.2.3 Timeline Extraction

Both Twitter streaming API and the Twitter search API do not allow the collection of data beyond a certain time period to be able to extract historical tweets. As a consequence, we collect our initial set of tweets within a fixed time window of 15 days. Because our goal was to study how the non-negotiable social identities of users correlated to their linguistic choices on Twitter, windowing the data by time period of 15 days could lead to high survivorship bias where users with higher activity within the chosen days could introduce bias in our analyses by having a higher influence. This is why, we decided to augment our data with timelines of identified individual users. This may not remove the survivorship bias completely, but may help mitigate it.

At the end of timeline extraction, we only retain one copy of each of the tweets. More concretely, to avoid over-inflating the effect of certain tweets that are more viral than the other, we use only unique tweet texts. This is an important preprocessing step to conduct a sociolinguistic

frequency-based analysis.

2.2.4 Data Statistics

At the end, our sociolinguistic analysis is conducted on an overall 6262 Twitter users with an aggregate of 588,110 tweets. This included 3295 pro-vaxxers with 310461 pro-vaccination tweets, and 2967 anti-vaxxers with 277649 anti-vaccination tweets, making it an average of about 94 tweets per user for both pro- and anti-vaxxers.

2.3 Methodology

We conduct two types of analyses to characterize the two competing groups: *linguistic analysis* and *network analysis*.

2.3.1 Linguistic Analysis

We test three linguistic variables which are described as follows.

Linguistic Intensification

We first study the differences in the usage of linguistic intensifiers. Intensifiers are words, or phrases that strengthen the meaning of other expressions and show emphasis. Examples include amplifiers (eg.“really”, “very”), usage of swear words, general interjections (eg. “wow”, “omg”), and exclamations. Intensifiers are commonly used to bolster argumentation to persuade the target audience. We hypothesize that users that are pro-vaxxers use more intensifiers. This is because pro-vaxxers have been found to frequently debunk anti-vaxxers' claims with scientific evidence [20]. Therefore, they would seem to take the corrective approach intended to persuade anti-vaxxers, hence using more intensifiers.

Pronominal Usage

Pronouns play a key role in models of narrative and discourse processing [46]. Because most of the vaccine-related misinformation is based on personal anecdotes, we would expect pronominal usage to be high amongst anti-vaxxers. To test this, we identify various different categories of pronouns (eg. “subject pronouns”, “object pronouns”, “third-person pronouns”), a complete list of which can be found in Table 2.2.

Use of Uncertainty Words

Previous research [43] has found the use of uncertainty words (eg. “might”, “likely”) as a negative linguistic correlate of echo-chamberness. This is based on the hypothesis that because users not in echo-chambers are exposed to alternate views, they may be less certain of their ideas. We adopted the list of uncertainty words from [43] to test if that is true i.e. if there is a significant difference in the use of uncertainty words across the two vaccination communities.

Table 2.2: This table shows the lexical categories we use for the sociolinguistic analysis along with the chosen list of words for each category (lexicon).

Lexical Category	Lexicon (vocabulary)
Intensifiers	
Amplifiers	*amazingly, -ass, astoundingly, awful, bare, bloody, crazy, dead, dreadfully, colossally, especially, exceptionally, excessively, extremely, extraordinary, fantastically, frightfully, fucking, fully, hella, holy, incredibly, insanely, mad, mightily, moderately, most, outrageously, phenomenally, precious, quite, radically, rather, real, really, remarkably, ridiculously, right, sick, so, somewhat, strikingly, super,supremely, surpassingly, terribly, terrifically, too, totally, uncommonly, unusually, veritable, very, wicked*
Swear words	*fu*****, etc.* A complete list of words can be found on Wikipedia's English swear words page [].
General interjections	*wow, hooray, ouch, uh oh, ew, aw, omg*
Exclamation	*!**
Uncertainty words	*may, might, perhaps, maybe/may-be, potentially, possibly, likely, probably, probable, possible, think, seem, believe, presume, would be, could be*
Pronouns	
Demonstrative	*this, that, these, those*
Possessive	*ours, mine, yours, theirs, his, hers*
Quantifier	*few, several, some, all, much, one, fewer, many, more, most, plenty, less, little, enough*
Reflexive	*myself, herself, ourselves, themselves, yourself, himself, itself, yourselves*
First-Person	*I, we, us, me, myself, my, mine, our, ours*
Second-Person	*you, yours, you're, your*
Third-Person	*he, she, theirs, themselves, them, her, him, his, himself, hers, herself, it, its, itself, they*
Gendered third-person	*he, she, her, him, his, himself, hers, herself*
Subject	*I, she, he, they, we, you, it*
Object	*me, us, them, him, you, her, it*
IT	*it, it's, its, itself*

2.3.2 Network Analysis

We also compute three network-level measures to characterize the network structure of the two target communities. We describe each of these measures in detail in their respective sections below.

Network Density

Network density is defined as the ratio of actual connections and potential connections [47]. Dense networks tend to "groupthink" [82] where conformity of ideas is highly valued and difference of opinions is discouraged.

EI Index

The EI (External-Internal) index was developed by Krackhardt and Stern in [60] as a measure of dominance of external over internal ties. More concretely, assuming two groups based on some attribute, one group defined as internal and the other as external, the EI index is computed as follows:

$$EI = \frac{EL - IL}{EL + IL} \tag{2.1}$$

where EL represents the number of external links and IL represents the number of internal links. EI index is a useful proxy for identifying echo-chamberness.

Echo-chamberness

To compare the echo-chamber effect in the two vaccination groups, we also directly compute the echo-chamberness of the two communities. We use the following definition of echo-chamberness as defined in ORA-PRO [11]: For a given network G, the echo-chamberness (EC) is defined as:

$$EC = (r * d)^{1/3} \tag{2.2}$$

where r is the reciprocity [91] of graph G or the ratio of bi-directional edges and the total number of edges in G, and d is the density of graph G.

2.3.3 Evaluation

Test Statistics

For each sub-category of the linguistic features in Table 2.2, we use two test statistics to compute the difference between the two groups. These are as follows:

1. The overall proportion of tweets that contain any of the words for a given lexical category (T_1)
2. The mean of the proportions of tweets of individual users containing any of the words for a given lexical category (T_2)

We use these test statistics to compute (i) the difference of proportions between the two groups, and (ii) the difference of means of proportions between the two groups.

The first test statistic regards each tweet independently. We use the second test statistic to account for differences in the linguistic choices of individual users.

Statistical Significance:

For the first statistic, we use a two-sample z-test for the difference of proportions (Z_1). For the second statistic, we use an independent z-test for the difference in means (Z_2). For all the tests, our $\alpha = 0.05$.

2.4 Results and Discussion

2.4.1 Linguistic Analysis

The summary of our linguistic analysis across all the lexical categories can be found in Table 2.3.

Table 2.3: This table shows the summary of our analyses across all the linguistic categories. The first column shows the lexical category. The second and third columns show the first test statistic as a percentage for pro-vaxxers and anti-vaxxers respectively. The fourth and fifth column display the z-score and p-value for the z-test for the difference of proportions. The sixth and seventh columns show the second test statistic as a mean percentage for pro-vaxxers and anti-vaxxers respectively. The eighth and ninth columns display the z-score and p-value for the independent z-test for the difference in means

Lexical Category	T_1 (Pro)	T_1 (Anti)	z-score (Z_1)	p-value (Z_1)	T_2 (Pro)	T_2 (Anti)	z-score (Z_2)	p-value (Z_2)
Intensifiers	45.90%	50.60%	-36.25	$< .001$	11.63%	14.96%	-6.59	$< .001$
Amplifiers	31.40%	37.10%	-45.32	$< .001$	10.91%	13.66%	-5.66	$< .001$
Swear words	4.0%	5.60%	-27.40	$< .001$	.57%	1.04%	-3.26	$< .001$
General interjections	17.50%	16.70%	7.89	$< .001$	.43%	.58%	-1.37	.17
Exclamation	1.10%	2.20%	-34.17	$< .001$	-	-	-	-
Uncertainty words	5.7%	7.0%	-20.84	$< .001$	4.12%	5.07%	-3.23	.001
Pronouns	55.80%	62.20%	-49.68	$< .001$	55.94%	61.83%	-7.38	$< .001$
Demonstrative	17.63%	20.91%	-31.84	$< .001$	18.61%	21.73%	-5.20	$< .001$
Posessive	1.30%	1.60%	-9.39	$< .001$	1.49%	1.67%	-.92	.36
Quantifier	15.3%	16.0%	-6.70	$< .001$	15.20%	16.83%	-3.06	.002
Reflexive	.80%	.86%	-2.26	.02	1.49%	.92%	3.43	$< .001$
First-Person	21.20%	23.44%	-20.67	$< .001$	20.96%	22.54%	-2.45	.01
Second-Person	16.40%	18.5%	-20.69	$< .001$	15.22%	16.47%	-2.23	.03
Third-Person	14.8%	20.9%	-60.51	$< .001$	14.29%	20.84%	-11.74	$< .001$
Gendered third-person	3.60%	5.60%	-36.84	$< .001$	3.15%	4.92%	-5.96	$< .001$
Subject	28.90%	37.50%	-69.53	$< .001$	27.64%	35.55%	-10.89	$< .001$
Object	21.64%	26.90%	-46.77	$< .001$	19.66%	24.51%	-7.91	$< .001$
IT	8.30%	10.29%	-26.16	$< .001$	8.21%	9.44%	-3.07	.002

Linguistic Intensification

We observe that our initial hypothesis that pro-vaxxers use more intensifiers is false. What we find is that anti-vaxxers employ significantly more linguistic intensifiers than pro-vaxxers. This holds true across all the sub-categories of intensifiers with the exception of the use of general interjections where the difference is marginal and not significant. While intensifiers are used as a persuasion technique, the observed results can possibly be explained by an old theory in speech communication that correlates the use of intensifiers with perceived powerlessness [21, 54]. Intensifiers and hedges are used more generally by people with low social power [21]. Because anti-vaxxers are a minority group, it is a possible argument one could make as perceived minority leads to perceived low social power which could lead to high linguistic intensification.

Pronominal Usage

From our analyses, we find that with the exception of reflexive and possessive pronouns, anti-vaxxers show a significantly high pronominal usage across all the categories. This difference is prominent specifically for third-person, gendered third-person, subject, and object pronouns. In sociolinguistic literature, pronouns are predominantly linked with narrative discourse structure. For example object pronouns such as “him” or “his” and gendered third-person pronouns “he” or “she” have a referential property, where their semantic interpretation is dependent on what they are referring to. Anaphoric references define objects already defined in the discourse [94] which

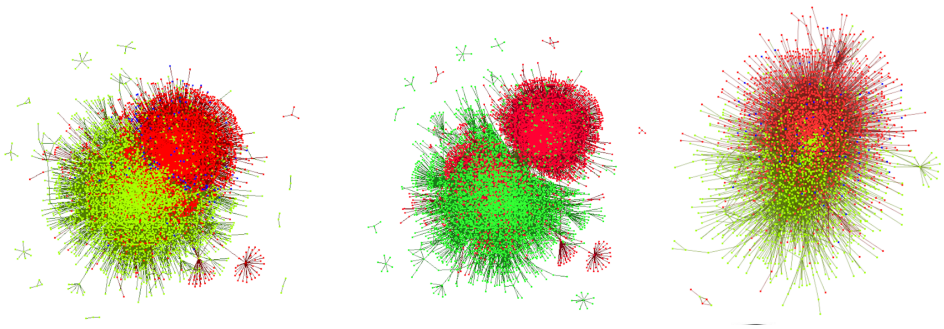
creates a better narrative viewpoint. Like intensifiers, pronouns are also found to be used heavily by people with lower levels of perceived power [69].

Use of Uncertainty Words

In terms of the use of uncertainty words, while we do find a significant difference between the two communities, we do not observe the same effect observed in the background literature [43]. In fact, we find a counter-intuitive result i.e. that the anti-vaccination community with higher echo-chamberness (as observed in section 2.4.2) tends to use more uncertain words than pro-vaccination community. This is an evidence that not all echo-chamber communities show certainty in their tweets as observed in [43].

2.4.2 Network Analysis

Figure 2.1: Mention (left), retweet (middle), and reply (right) networks of pro (in green) and anti (in red) vaccination communities created using ORA-PRO [10, 29]



Along with the linguistic analysis, we also compute various network level measures on the communication networks of the two target groups. These measures include the *network density*, *EI index*, and *echo-chamberness*. We also visualize the three communication networks as shown in figure 2.1. All the network-based measures, and graphs were computed using ORA-PRO [10, 29].

We observe that anti-vaccination communities tend to have higher network density, negative EI indices with higher absolute values, and higher echo-chamberness across all the communication networks. On the other hand, the EI index for the pro-vaccination communities is positive for mention and retweet networks displaying dominance of external ties. A summary of network-level measures can be found in Table 2.4. Interestingly from the network graphs we can observe that on some level the two competing groups are almost detached. This is specifically visible in the retweet network graph in Figure 2.1.

Furthermore, in the past researchers have argued in [50, 83] that anti-vaccination beliefs may be represented over a continuum as there are many reasons to not vaccinate, and hence, anti-vaccination community may be a combination of multiple sub-communities. On the other hand, it is also observed that pro-vaxxers are largely a unified group dominated by the established medical community [1]. In the light of the above, our network analysis results are, therefore,

surprising, as we observe anti-vaxxers to have a higher density of connections with more echo-chamberness.

Table 2.4: This table shows the network-level measures for the three types of networks: mention network, retweet network, and reply network

Measure	Mention Network	Retweet Network	Reply Network
Network Density	1.7e-5	1.1e-5	3.1e-6
Network Density (Pro)	1.5e-5	1.0e-5	2.2e-6
Network Density (Anti)	4.1e-5	3.2e-5	6.3e-6
EI Index (Pro)	0.025	0.023	-0.167
EI Index (Anti)	-0.276	-0.432	-0.572
Echo-chamberness (Pro)	0.0064334823	0.005364444	0.0043579605
Echo-chamberness (Anti)	0.009268834	0.007850341	0.005905038

2.5 Limitations and Future work

One minor limitation of our study is that in the data collection phase, the number of collected hashtags for the two communities was unbalanced. This could potentially have introduced some bias in our downstream tasks such as label propagation. A possible limitation pertaining to the network analysis is that we do not normalize our EI indices to avoid losing precision. This, however gives us stronger results as while the nodes in the anti-vaccination network are lower than the pro-vaccination network, the EI index for anti-vaxxers is more negative than pro-vaxxers.

Another limitation of our study is that we make a crude and simplifying assumption that the stance of users within the vaccination discourse is binary (pro or anti). As discussed in [50, 83], this may not be the case as anti-vaccination beliefs may be represented over a spectrum. In the future, we intend to extend our analysis over different sub-communities defined over varying levels of vaccine-related hesitancy.

A limitation pertaining to our choice of seed hashtags for the label propagation algorithm is that we assume that there is no overlap in the use of hashtags between the two groups of users. While this is possible, there are three factors that suggest this is not the case: (i) Our hashtag validation approach showed that 97.5% of the pro-vaccination hashtags retrieved pro-vaccination tweets, and 90.5% of the anti-vaccination hashtags retrieved anti-vaccination tweets. This suggests that the tweets that contained both pro- and anti-vaccination hashtags were fewer; (ii) our network analyses showed that members of groups identified by label propagation often retweeted each other; and (iii) validation of our label propagation algorithm showed that 96% of the labeled pro-vaxxers were pro-vaxxers, and 80%of the labeled anti-vaxxers were anti-vaxxers.

Finally, all our analyses are correlational in nature, and do not depict causation. This remains to be one of the important future directions to test whether a certain network characteristic causes linguistic changes in the network or vice-versa.

2.6 Conclusion

In this part, we have carried out a comparison between two online competing vaccination communities: *pro-vaxxers* and *anti-vaxxers*. We have studied these communities in relation to their linguistic and social interactions. We conduct two kinds of analyses: (i) linguistic, and (ii) network-level. We observe anti-vaxxers to display more frequent usage of linguistic intensification, pronouns, and uncertainty words. We also observe significant differences in the network structures of the two communities with *anti-vaxxers* displaying higher echo-chamberness. These results suggest that anti-vaxxers form a tighter community prone to the presentations of anecdotes, and so may be more resistant to factual knowledge from outside the group.

Part II

Exploring COVID-19 Misinformation

Chapter 3

Characterizing COVID-19 Misinformation Communities Using a Novel Twitter Dataset

From *conspiracy theories* to *fake cures* and *fake treatments*, COVID-19 has become a hot-bed for the spread of misinformation online. Social media is known to facilitate manipulation and radicalization of users. Since public actions affect the public health and safety directly, it is more important than ever to identify methods to debunk and correct false information online. In this part of the thesis, we conduct analyses to characterize the two competing COVID-19 communities online: (i) *misinformed users* or users who are actively posting misinformation in the form of fake cures, conspiracies, false preventions, and fake treatments, and (ii) *informed users* or users who are actively spreading true prevention or calling out and correcting misinformation. The goals of this study were two-fold: (i) collecting a diverse set of annotated COVID-19 Twitter dataset that can be used by the research community to conduct meaningful analysis; and (ii) characterizing the two target communities in terms of their network structure, linguistic patterns, and their membership in other communities online.

It is important to note that unlike Chapter 1, there is not an overarching single issue related to COVID-19 to be pro or anti, but in fact a number of issues. Therefore, while the strategy to understand the discourse is similar, we define the communities around their membership within informed and misinformed tweets. Misinformation around COVID-19 can come in many forms and issues, and each of the issues can have a pro and anti community around it. The first step to understand misinformation is to understand the types of misinformation, and whether they can be accurately identified based on the features associated to them, and this chapter lays a groundwork for that.[1]

In this part, we first describe the background literature relevant to the different available COVID-19 datasets, and also relevant to different types of analyses that the research community has conducted. We then describe our novel dataset *CMU-MisCOV19* containing 17 different categories, with 4573 annotated tweets. Finally, we conduct network analysis to compare the

[1]This work was accepted at the 5th International Workshop on Mining Actionable Insights from Social Networks (MAISoN) at CIKM 2020, a preprint of which can be found at [65]

two target communities in terms of their network density, bot analysis to identify the proportion of users depicting bot-like behavior, sociolinguistic analysis to understand the linguistic patterns of the two target communities to understand their traits and behaviors, and finally we explore the interplay of vaccination-related misinformation by identifying vaccination membership of the misinformed users.

3.1 Background

3.1.1 COVID-19 Datasets

In the short amount of time, many COVID-19 datasets have been released. Most of these datasets are generic, and lack annotations or labels. Examples include multilingual corpus on a wide variety of topics related to COVID-19 [5, 33, 56], longitudinal Twitter chatter dataset [15], multilingual dataset with location information of the users [73], Twitter dataset for Arabic tweets [8], Twitter dataset for popular Arabic tweets [51], and dataset for identification of stance, replies, and quotes [89]. Most of these datasets either have no annotations at all, employ automated annotations using transfer learning or semi-supervised methods, or are not specifically designed for misinformation.

In terms of datasets collected for COVID-19 misinformation analysis and detection, examples include CoAID [36] which contains automatic annotations for tweets, replies, and claims for fake news; ReCOVery [98] is a multimodal dataset annotated for tweets sharing reliable versus unreliable news, annotated via distant supervision; FakeCovid [77] is a multilingual cross-domain fake news detection dataset with manual annotations; and [37] is a large-scale Twitter dataset also focused on fake news. A survey of the different COVID-19 datasets can be found in [61] and [81].

In terms of the diversity of the classes, and the size of the dataset, the most relevant dataset is by Alam et al. [6] who, like our study, present a comprehensive codebook to annotate tweets on a finer granularity. Their dataset, however, is limited to a few hundred tweets, and our dataset is much more diverse in the range of topics covered. Dharawat et al. [39] present a similar dataset with focus on the severity of the misinformation. However, their dataset does not consider the different "types" of misinformation. Finally, Song et al. present a dataset in [84] which contains a diverse set of 10 categories, but still is not as large, and contains fewer categories in relation to the dataset collected within our study.

3.1.2 Misinformation Analysis

A plethora of research has already been conducted for analysing COVID-19 misinformation online. Some examples include categorization and identification of misinformed users based on their home countries, social identities, and political affiliation [55], [79], characterization of different types conspiracy theories propagated by Twitter bots [45], characterization of the prevalence of low-credibility information related to COVID-19 [93], exploratory analysis of the content of COVID-19 tweets [70, 78], understanding the types, sources, and claims of COVID-19 misinformation [22], and comparison of the credibility of COVID-19 tweets to datasets pertain-

ing to other health issues [26]. To the best of our knowledge none of the studies have characterized COVID-19 misinformation communities in terms of their sociolinguistic patterns. In this study, we do not characterize the misinformation *content* directly. Instead, we conduct a set of analysis to understand and characterize these *communities* through their content, and content-sharing behaviors and interactions.

3.2 Methodology

3.2.1 Data Collection

To collect Twitter dataset, we use the Twitter search API. For a given set of keywords, the search API usually only serves tweets from the past week. We use the search API to collect data in four phases. In the first three phases, we collect the data on three days: 29th March 2020, 15th June 2020, and 24th June 2020. Each of these collections extracted a set of tweets from their corresponding week. The division of data into phases was done to increase the diversity of topics covered, and to reduce selection bias. With each iteration, we also updated our list of hashtags to account for new themes. The complete set of keywords and hashtags used is in table 3.1. Because our goal is to characterize communities and their behaviors, in the fourth phase, we collected the timelines of users to augment our data. This was done using the search API on 27th July 2020. This step was done to increase per-user data for annotation and analysis. In the end, the earliest tweet in our *annotated* dataset is from 9th January 2020, and the latest tweet is from 8th July 2020.

Table 3.1: This table shows the hashtags, and keywords we used in conjunction with "coronavirus" and "covid" to collect data from Twitter

Type	Terms
Keywords	*bleach, vaccine, acetic acid, steroids, essential oil, saltwater, ethanol, children, kids, garlic, alcohol, chlorine, sesame oil, conspiracy, 5G, cure, colloidal silver, dryer, bioweapon, cocaine, hydroxychloroquine, chloroquine, gates, immune, poison, fake, treat, doctor, senna makki, senna tea*
Hashtags	*#nCoV20199, #CoronaOutbreak, #CoronaVirus, #CoronavirusCoverup, #CoronavirusOutbreak, #COVID19, #Coronavirus, #WuhanCoronavirus, #coronaviris, #Wuhan*

3.2.2 Data Annotation

Our annotation task aims to determine the category to which a given tweet belongs to. After many discussions and revisions, we identify 17 categories that a particular tweet could classify to. These 17 categories are defined in table 3.2. These categories are defined in further detail along with their definitions and examples in our codebook which we make available for the public to use.

Table 3.2: This table describes the categories we identified to classify/annotate tweets along with the distribution of annotations as identified by Annotator 1 in the first phase.

Category	Count
Irrelevant	131
Conspiracy	924
True Treatment	0
True Prevention	175
Fake Cure	141
Fake Treatment	34
False Fact or Prevention	321
Correction/Calling out	1331
Sarcasm/Satire	476
True Public Health Response	163
False Public Health Response	3
Politics	512
Ambiguous/Difficult to Classify	143
Commercial Activity or Promotion	37
Emergency Response	17
News	95
Panic Buying	70

Based on these categories, tweets were randomly and uniformly sampled from the data collection to maintain diversity in terms of topics covered. In the first phase 4573 tweets were annotated by a single annotator. Table 3.2 shows the distribution of the data in terms of the different categories as annotated by the first annotator. In the second phase, 651 of these annotated tweets were assigned randomly to 6 other annotators.

3.3 Data Description

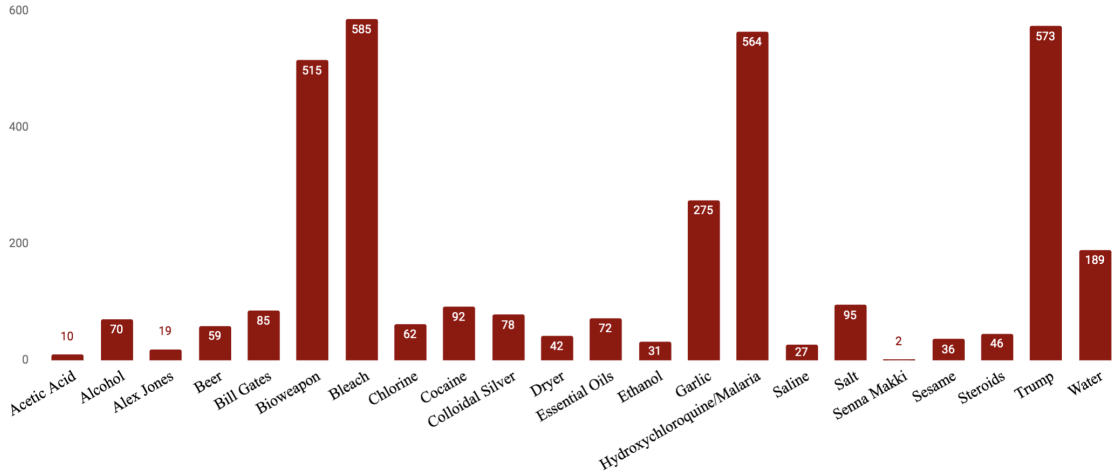
Our data collection strategy is different from others in two main aspects: (i) we have a diverse set of categories taking into consideration different types of information and misinformation online; and (ii) our dataset is one of the very few, if not the only one, with emphasis on informed communities with categories such as "True Prevention", "Calling out/correction", "True Public Health Response", and "Sarcasm". We believe this is necessary as building models requires not just the annotation of false information, but as well as complementary true information categories.

At the end, we have 4573 annotated tweets, comprising of 3629 users with an average of 1.24 tweets per user. Our annotated data not only covers a wide range of categories as observed in table 3.2, but also covers a wide range of topics as can be seen in figure 3.1. We call this dataset *CMU-MisCOV19* [64].

In adherence to the FAIR principles, the database and the codebook has been uploaded to Zenodo and is accessible with the following link: http://doi.org/10.5281/zenodo.4024154. In adherence to the Twitter's terms and conditions, we do not provide the full tweet JSONs, but provide the tweet IDs so that the tweets can be rehydrated. We also provide the annotations,

and the date of creation for each tweet for the reproduction of the results of our analyses. The annotated tweets are included in a CSV file with the following fields: *status_id* (tweet id of the tweet), *status_created_at* (timestamp of the creation of the tweet), *annotation1* (annotated class of the tweet by the first annotator), and *annotation2* (annotated class of the tweet by the second annotator, if exists).

Figure 3.1: This chart shows the frequency of each identified topic across all the tweets. Note: Some tweets may have more than one topic.



3.4 Analysis and Discussion

3.4.1 Identifying Communities

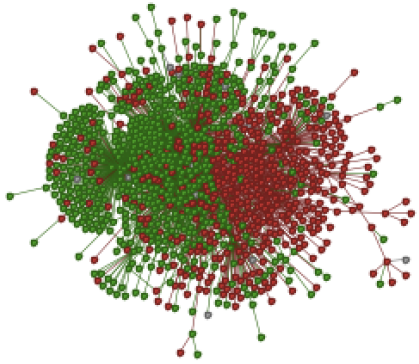
Conducting analyses for a competing set of communities requires identifying those communities first. Because we have already annotated data across a set of true and false information categories, we identify the membership of the users by assigning a valence of +1 to the categories *True Treatment, True Prevention, Correction/Calling Out, Sarcasm/Satire, and True Public Health Response*, and a valence of -1 to the categories *Conspiracy, Fake Cure, Fake Treatment, False Fact or Prevention, and False Public Health Response*. Note that we assign the valence to the categories (or annotations) and not the tweets themselves. This is so that we can leverage the annotations from multiple annotators. At the end, we compute the valence of each user as a weighted sum of the valence of the annotations assigned to their tweets. Then we use the valence assigned to each user to identify their membership i.e. if valence is greater than 0, the user is assigned to the *informed* group, and if the valence is less than 0, the user is assigned to the *misinformed* group. Out of 3629 users, the community detection process assigns 47% (1697) of the users to the informed group, 29% (1043) of the users to the misinformed group, and 24% (889) of the users to ambiguous or irrelevant category[2].

[2]Irrelevant users are users who have only posted tweets within other categories such as “Politics” or “Emergency Response”. Because these categories do not have an assigned valence related to misinformation, they are not relevant for the purposes of this study.

3.4.2 Network Analysis

To conduct network analysis, we first extract only the COVID-19 related tweets from the timelines of each user. We do this by filtering all the tweets by the case-insensitive keywords *"corona"* and *"covid"*. We then extract the retweet, mention, and reply networks of the two target communities, and combine those networks together. We then compute the *network density* for each of the two groups. As described in [66], network density is defined as the ratio of actual connections and potential connections. In dense networks, conformity of the ideas is highly encouraged, and difference of opinions is discouraged. We also use ORA-PRO [9, 10, 30] to plot the network graph as shown in figure 3.2

Figure 3.2: Retweet+Mention+Reply network with informed users (in green) and misinformed users (in red) created using ORA-PRO [10, 29]. Note: Users with ambiguous or unidentified membership have been removed from the graph for simplicity.



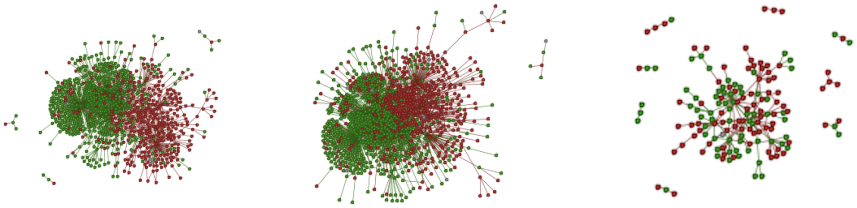
We note that both the informed and misinformed users display echo-chamberness with misinformed sub-communities being much denser than the informed sub-communities as shown in table 3.3. We do, however, notice some two-way communication from both sides.

Table 3.3: This table shows the number of nodes, links, and the network density for the two target sub-communities.

Measure	Overall	Informed	Misinformed
Nodes	2477	1515	923
Links	2947	1489	826
Network Density	4.8e-4	6.5e-4	9.7e-4

We also plot the retweet, mention and reply network separately as shown in figure 3.3. While retweet, and mention network show little to no two-way communication, we can observe that the reply network, while small in size, does in fact have much more inter-group engagement. We hypothesize that this is likely a consequence of the "corrective" or "calling-out" behavior.

Figure 3.3: Retweet (left), mention (middle), and reply network (right) with informed users (in green) and misinformed users (in red) created using ORA-PRO. [10, 29]



3.4.3 Bot Detection

To understand the role of bots within the two competing groups, we used Bot-Hunter [16, 17, 18, 19] which has a precision of .957 and a recall of .704, to identify potential bot-like accounts. We use the probability of greater than or equal to .75 as our confidence threshold to identify bots. We use a two-sample z-test for the difference of proportions ($\alpha = 0.05$) to test the difference in proportion of bots between the two competing groups of users. The results of our analyses can be found in table 3.4.

Table 3.4: This table shows the number and percentage of bots within each of the two competing groups

Measure	Overall	Informed	Misinformed
Number of Users	3629	1697	1043
Number of Bots	505	184	202
Percentage of Bots	14%	11%	19%

We observe that from a total of 3629 users, 14% (505) of the users are identified as bots. The percentage of bots within identified misinformed users, however, is much higher (19%) than within identified informed users (11%). We find our results to be statistically significant ($p < 0.001$; $z = -6.23$). This indicates that more than 1/5th of the misinformation related posts in our dataset are potentially result of disinformation campaigns related to COVID-19.

3.4.4 Sociolinguistic Analysis

To understand the linguistic differences between the two competing communities, we conduct a linguistic analysis based on the tweets of the two groups by using the Linguistic Inquiry and Word Count (LIWC) program [72]. LIWC is a text analysis tool which looks at the different lexical categories each of which is psychologically meaningful. For a given text, LIWC calculates the percentage of each LIWC categories. All of these categories are based on word counts.

We run the LIWC program on the timelines of all the members for each of the two competing groups. We only use tweets relevant to COVID-19. We also remove users identified as bots. Because some users may be more active than others, using the results of the program as is may introduce biases in our analyses. To account for those biases, we first normalize the percentages

by the size of the data for each user. We use the mean of the normalized LIWC indices of tweets of individual users for a given lexical category as our test statistic. We use an independent z-test for the difference in means to establish statistical significance. For all our tests, $\alpha = 0.05$. Our analyses are summarized in table 3.5.

Table 3.5: This table shows the summary of our analyses across all the linguistic dimensions described above using LIWC. The first column shows the lexical category. The second and third columns show the test statistic (M_1) as the mean of the LIWC indices for informed and misinformed communities respectively. The fourth and fifth columns display the z-score and p-value for the independent z-test for the difference in means.

Lexical Category	M_1 (Informed)	M_1 (Misinformed)	z-score (Z_1)	p-value (Z_1)
function	33.90	29.32	7.25	$< .001$
tentat	1.87	1.63	1.66	.1
certain	1.14	1.32	-1.57	.1
pronoun	7.97	6.53	4.89	$< .001$
ipron	3.26	3.03	1.23	.2
ppron	4.71	3.49	5.39	$< .001$
Analytic	69.83	76.01	-4.82	$< .001$
social	6.49	5.05	5.45	$< .001$
family	.34	.20	2.24	.03
friend	.17	.17	-.03	.97
Authentic	25.12	16.43	6.78	$< .001$
Tone	35.42	37.59	-1.45	.15
informal	4.89	5.16	-1.63	.10
swear	.51	.34	1.86	.06

For this part, we focus on investigating four linguistic dimensions, each of which, along with its linguistic correlates, is described below.

Narrative Discourse Structure

Narratives play a central role in how individuals process information, communicate, and reason [88]. We set to test the differences in the usage of narratives or anecdotes between the two COVID-19 misinformation communities. The LIWC correlates for narrative discourse structure include high use of function words, pronouns, analytic summary dimension, and authenticity. High usage of function words and pronouns happens more often when expressing feelings and behaviors which tends to happen frequently in narratives [71]. Moreover, low analytical thinking also suggests narrative language [72]. Furthermore, authentic individuals tend to be more personal, humble, and vulnerable [72]. Therefore, we use all of these as proxies to identify variation in the use of narratives across communities.

In the past [66], it has also been suggested that misinformed communities (eg. anti-vaxxers) tend to use many more pronouns suggesting highly narrative discourse structure. In this anal-

ysis, however, we find that informed users in the COVID-19 discourse use significantly more pronouns, more functional words, mention more family-related keywords, are less analytical, and more authentic and honest in comparison to misinformed users. All of these suggest that informed users use many more narratives than misinformed users. This is an interesting finding as it presents a dichotomy between the different misinformation communities (eg. anti-vaxxers and COVID-19 misinformed community). In hindsight, this is also an intuitive result, as our informed group is obtained from corrective discourse where users present their stories of family members or friends suffering from COVID-19 to call out conspiracies and false information. Because the two communities still seem to have less two-way communication, this also suggests that just the content and framing of the message may not be enough, and perhaps there is a need to connect the two groups by identifying an effective medium.

Tone

Tone describes how positive a given text is. According to the definition by LIWC, the higher the LIWC index for tone, the more positive the tone. Index values less than 50 typically suggest a more negative tone. While we do not see significant differences in the emotional tone of the competing groups, we find both the communities to be highly negative.

Linguistic formality

Formality of the language has often been considered as one of the most important dimensions for stylistic variation. In [48], authors define linguistic formality as a style of writing that is meant to be precise, coherent, articulate and convincing to an educated audience, as opposed to informal discourse which is filled with deictic references (eg. here, there), pronouns, and narration. The LIWC correlates to this dimension are swear words (swear), and informal language (informal). Informal language in LIWC is computed on the bases of swear words, netspeak (eg. btw, lol), nonfluencies (eg. err, hmm), assents (eg. agree, OK), and fillers (eg. youknow).

From table 3.5, it can be observed that misinformed users tend to be more informal than informed users, though informed users tend to use more swear words than misinformed users. This is intuitive as many of our informed users post corrective or sarcastic tweets to call out misinformation. However, our results are not significant, and, hence inconclusive.

Linguistic Uncertainty

It has been suggested in the past studies[43] that dense communities with high echo-chamberness, tend to be more certain as there are less differing opinions. We use LIWC's "certainty", and "tentative" categories as correlates to identify linguistic uncertainty. While we do find denser communities – i.e. misinformed users – to be more certain, our results are not significant, and, hence, inconclusive.

3.4.5 Vaccination Stance

To understand the interplay between the different kinds of misinformation themes and communities, we identify the vaccination-related stance of the members of the misinformed sub-

community. To do that, we first identify the subset of misinformed community who have posted at least one tweet related to "vaccines" in the past. We then collect the user-to-hashtag co-occurrence network. We use the valence of the vaccination hashtags obtained in part 1 to identify the stance of each member (pro vs. anti) based on the weighted sum of the valences. If the weighted sum is greater than 0, we identify the member as pro-vaxxer, and if the weighted sum is less than 0, we identify the member as anti-vaxxer. The distribution of the pro- and anti-vaxxers within the COVID-19 misinformed group is as shown in table 3.6.

Table 3.6: This table shows the number and percentage of pro- and anti-vaxxers within the misinformed group.

Measure	Value
Users w/ vaccine-related tweets	2750 (out of 3629)
Misinformed users	1027 (37%)
Anti-vaxxers	423 (41%)
Pro-vaxxers	224 (22%)
Ambiguous	380 (37%)
Misinformed pro-vaxxer bots	37 (17%)
Misinformed anti-vaxxer bots	82 (22%)

We observe that from 1027 COVID-19 misinformed users in our dataset, 41% of the members are identified as anti-vaxxers, whereas only 22% of the members are identified as pro-vaxxers. The difference between the proportions of the two communities is striking. We also identify the proportion of bots within each of the two groups: *misinformed pro-vaxxers*, and *misinformed anti-vaxxers*. As shown in table 3.6, 17% of the misinformed pro-vaxxers are bots, which is significantly lower than the proportion of bots within the misinformed anti-vaxxers. The first thing this suggests is that a big chunk of COVID-19 misinformation online may in fact be *disinformation*, and hence, intentional. The existence of bots within both the informed and misinformed communities also suggests that much of the disinformation online may be an organized effort to amplify the COVID-19 debate to create discord in the communities as seen in the past with Twitter bots and Russian trolls [25].

3.5 Limitations

The first important limitation pertaining to our work is that most of our analyses are based on the data that has been annotated by only 1 annotator. We try to mitigate this by having more than 1/7th of our annotations annotated by a second annotator, and taking into account all those annotations while computing the membership of each user. Another limitation to our work is that all our analyses are correlational in nature, and do not depict causation.

A limitation pertaining to our data collection strategy is that we collect our data across a period of three weeks, augment our data with timelines of users, and update our list of hashtags to account for new themes. We then sample a subset of this data for annotation process. Because of the way data was collected, it cannot be used for assessing change over time. Moreover, while

this ensures the diversity of misinformation-related topics and agents, it may limit our ability to estimate the actual extent to which the different types of stories are more or less present.

Finally, one minor limitation related to our bot analysis is that we use a second-level inference from a trained model. We try to mitigate this by using labels with probability greater than or equal to .75 to ensure high quality labels.

3.6 Conclusion

In this part, we present a methodology to characterize the competing COVID-19 misinformation communities by comparing them in terms of their network structure, sociolinguistic variation, and membership in disinformation campaigns and in other health-related misinformation communities such as anti-vaxxers. We find that even though COVID-19 is a recent event, misinformation related to it has created a set of polarized communities with high echo-chamberness. Misinformed communities are observed to be denser than informed communities which is in line with previous studies such as [66]. We find that bots exist in both the informed and misinformed groups, but the percentage of bots in misinformed users is significantly higher suggesting the prevalence of disinformation campaigns. Our sociolinguistic analysis suggests that both the target communities depict negative emotional tone in their posts, with signals that informed users use many more narratives than misinformed users. Finally, we discover that many misinformed users may be anti-vaxxers. Our analyses suggest that misinformation communities are much more complex as they are highly organized, and tend to be highly analytical. Unlike previous suggestions [76], they may not be responsive to narrative correctives, and hence, a "one size fits all" generic messaging intervention for debunking misinformation may not be a feasible solution. A successful intervention may require to identify, and ban the disinformation campaigns. It may also be useful to identify the right medium of communication to connect the two groups. This can be achieved by identifying users in misinformed communities who are not *rebroadcasting*, or have high betweenness centrality to be messengers for disseminating factual information. It may also be useful to further understand the linguistic patterns and preferences of these communities to create an effective *content* and *framing* of the messaging.

Chapter 4

Conclusion and Future Work

In this thesis, we explore the characteristics of the two online health-based misinformation communities: *vaccination communities* and *COVID-19 misinformation communities*. We characterize them by juxtaposing them alongside their competing groups to understand their interactions and linguistic patterns. While both of these are health-based communities, there are some differences between them: First, vaccine hesitant groups have existed for as long as vaccines themselves [20]. Anti-vaccination beliefs, therefore, have deeper roots, and have persisted through decades. On the other hand, COVID-19-related communities are *event-based*, relatively new, and have arguably short life. Second, controversy around vaccination communities revolves mostly around childhood vaccinations with concerned parents. COVID-19, on the other hand, is believed to not affect children as much as adults [23]. This means the demographics of the two communities may be different. Finally, there is an element of *immediacy* to the COVID-19 crisis, exerting a lot of pressure on the community at large, and hence the misinformation communities as well. This is also because public actions have an immediate impact on public health. The public health response to the anti-vaccination campaigns, on the other hand, is relatively slow. Despite these differences, we observe similar network interaction patterns across the two communities, and their corresponding misinformed sub-communities. We do observe some differences across linguistic patterns though. One prominent and interesting difference is the nature of narrative discourse structure. In the vaccination-related communities, we observe misinformed community displaying more signs of narrative discourse. This is in line with the theory that misinformation thrives on false narratives. In the COVID-19 misinformation communities, however, we find convincing evidence that informed community uses many more narratives and anecdotes than misinformed community. This is a result of the changes in discourse between the two topics, and in the causality of events: In vaccination discourse, misinformed users share personal stories to convince the world of the harms of vaccines. In COVID-19 discourse the social burden of responsibility shifts to informed users who share personal stories to call out those who think COVID-19 is a hoax and is not real.

Narratives are generally considered to be tools of persuasion [38]. Many believe that narrative messaging can be used to combat misinformation [31, 76, 80]. While the framing of the message is important, from the analysis in part 2, we observe that it is not enough. There needs to be a mechanism for the informed and misinformed communities to have a two-way communication. In other words, it is important to identify the method of dissemination of information. Identifying

individuals with high betweenness centrality as messengers would be one way to achieve that. But, without identifying the right medium of information dissemination, the informed communities, due to high echo-chamberness, will likely have no effect on the misinformed communities.

Finally, it is also very important to identify disinformation campaigns and ban the bots and trolls involved in spreading misinformation. This requires building sophisticated models to automatically identify different types of misinformation. Because these models are typically supervised "data-hungry" machine learning models, they require large-scale annotated databases. In this thesis, we present a relatively large dataset along with a comprehensive codebook which we make public as a "call to arms" for the community to use and collect more data.

In terms of future work, there are various directions one can take. One promising line of work would be in the direction of actively testing message interventions on online platforms using linguistic cues such as those described in our work. A relevant work in this regard is by Munger [68] on the experimental reduction of racist harassment on Twitter. It may also be important to *simulate* these communities to identify the right medium to foster engagement across communities.

Another interesting direction would be to identify a systematic approach of characterizing communities both in terms of stance and disinformation. Our work suggests that linguistic cues may be helpful in defining stance of users. Future work should further explore the relation between stance and disinformation vis-a-vis issues. Many topics, such as COVID-19, do not have a clearly defined overall stance. This is because there are many sub-topics associated with them, each of which can have its own stance. For example, within the topic of COVID-19, relevant subtopics include COVID-19 vaccination, COVID-19 government response, and wearing masks, each of which can have a pro and anti stance. Our codebook in Appendix A identifies many types of misinformation themes and these cut across the stance on these different issues. In the future work, it is important to look at categorizing the nature of the problem by focusing on stance and misinformation together within the study of online discourse.

Another interesting direction would be to look at the *dynamics* of the different communities to understand how these communities are formed and dissolved over time, and what offline *events* play significant role in those changes.

Finally, an obvious next direction based on the second part of this thesis would be to use the dataset we have collected and annotated to build models for the detection of different types of misinformation, and to use the codebook we have created (in Appendix A) as a basis to collect more annotations.

Appendix A

CMU-MisCov19 Codebook

This chapter contains the codebook we created for collecting annotations for *CMU-MisCov19*, a COVID-19 Twitter Dataset for Misinformation. This codebook was created after multiple discussions and revisions, and the final dataset was corrected to reflect this version of the codebook.

A.1 Coding Scheme

Table 1 shows the list of categories we will consider for annotating Tweets related to COVID-19 misinformation on Twitter.

Table A.1: This table describes the categories along with their IDs

ID	Category
0	Irrelevant
1	Conspiracy
2	True Treatment
3	True Fact or Prevention
4	Fake Cure
5	Fake Treatment
6	False Fact or Prevention
7	Correction/Calling out
8	Sarcasm/Satire
9	True Public Health Response
10	False Public Health Response
11	Politics
12	Ambiguous/Difficult to Classify
13	Commercial Activity or Promotion
14	Emergency Response
15	News
16	Panic Buying

A.2 Description

The following section describes each of the categories in detail, along with their corresponding examples and justifications.

0. Irrelevant

A tweet shall be classified as irrelevant if it may or may not mention COVID-19 or SARS-Cov-2, but if it cannot be classified in *any* of the other categories below.

Example(s):

Tweet	**Justification**
"If you're feeling like it, today is "Update Friday" so dip into a channel relevant to your interests and answer a question or pose one. Let's get chatty."	This tweet should be marked as irrelevant as it is not relevant to COVID-19 or SARS-Cov-2.
"If Taylor Swift and Avril Lavigne collaborate on a song together it would be the cure for coronavirus. https://t.co/90NxHkmJcc"	This tweet mentions *Coronavirus*, but ultimately is unrelated in terms of content, and hence should be marked as irrelevant.
"The cure against Coronavirus is 'Kings &; queens' by @avamax stream it now! https://t.co/idoU6gRLHN"	This tweet mentions *Coronavirus*, but ultimately is unrelated in terms of content, and hence should be marked as irrelevant.
"@twlldun It's not fair I was into Covid before anyone heard of it. Now everyone is all like 'OMG Covid 19' like they invented it. Where were you when it was Covid 1-18 guys? It's Bleach all over again it's not fair. Am I doing this right?"	This tweet mentions *COVID-19*, but ultimately is unrelated in terms of the content, and hence should be marked as irrelevant.

1. Conspiracy

A tweet shall be classified as a conspiracy if it *endorses a conspiracy story*. Some examples of conspiracy themes related to COVID-19 include:

1. It is a *bioweapon*.
2. Electromagnetic fields and the introduction of *5G wireless technologies* led to COVID-19 outbreaks.
3. This was a plan from *Gates Foundation* to increase the Gates' wealth.
4. It leaked from the *Wuhan Labs* or Wuhan Institute of Virology in China.
5. It was *predicted* by Dean Koontz.

Examples:

Tweet	Justification
"Interesting interview with Prof. Frances Boyle re. #COVID19 is indeed a perfect #Bioweapon! Smoking gun proof! Scary shit!"	This tweet shall be marked as a conspiracy as it is endorsing Prof. Frances Boyle's video on COVID-19 being a bioweapon.
"Starting to blame US for Covid-19. This is ridiculous. It came from wuhan. CIA has stated in leaked documents that the wuhan laboratory is, in fact, a "hidden" bioweapon engineering facility. The only lab in all of China rated high enough for handling such diseases. It probably-"	This tweet shall be marked as a conspiracy as it endorsing the misinformation that COVID-19 was leaked from a lab in Wuhan.
"i think both 5G and covid-19 lowering immune system. considering the great suspicion that covid-19 is actually an offensive warfare bioweapon, they maybe designed to work together, greatly increasing coronavirus lethality. turning 5G off might be the antidote to coronavirus."	This tweet shall be marked as a conspiracy as it is endorsing the misinformation that 5G is responsible for the COVID-19 outbreak, and that 5G weakens the immune system.

2. True Treatment

A tweet shall be classified as a true treatment if it endorses a method of treatment to ease the pain (rest and sleep, keep warm, drink plenty of liquids, etc.), and if any of the following conditions are met:

1. The treatment has been *verified by the World Health Organization (WHO)* site.
2. The treatment has been *verified by the Center of Disease Control and Prevention (CDC)* site.
3. The treatment is *supported by a peer-reviewed scientific journal* that appears in Ulrich's Global Serials Directory as both "Active" and "Refereed/Peer-reviewed".
4. The treatment is *supported by a publicly posted working paper* authored or co-authored by tenure track faculty at a university in the top-800 universities worldwide according to the Times Higher Education World University Rankings 2019.
5. Tweet links directly to news story which *correctly* cites a peer-reviewed journal article (using standards above to adjudicate the journal).

Example:

Tweet	Justification
"Mild fever itchy throat doc says sleep well drink fluids (the non alcoholic kind he stressed). I feel anxious about covid then remember I have these symptoms 4 times every year."	This tweet shall be marked as a true treatment as the tweet endorses some of the treatments for self care verified by WHO.

3. True Prevention

A tweet shall also be classified in this category if it explicitly endorses a method of prevention and any of the following conditions are met:

1. The prevention has been *verified by the World Health Organization (WHO)* site.
2. The prevention has been *verified by the Center of Disease Control and Prevention (CDC)* site.
3. The prevention is *supported by a peer-reviewed scientific journal* that appears in Ulrich's Global Serials Directory as both "Active" and "Refereed/Peer-reviewed".
4. The prevention is *supported by a publicly posted working paper* authored or co-authored by tenure track faculty at a university in the top-800 universities worldwide according to the Times Higher Education World University Rankings 2019.

Some examples of the known true preventions of the COVID-19 disease from the CDC site include:

1. Washing your hands often
2. Avoiding close contact
3. Covering your mouth and nose
4. Covering coughs and sneezes
5. Cleaning and disinfecting
6. Monitoring your health

Note: A tweet "No, cocaine does not prevent coronavirus" would not fall into this category as while it may be preventative (i.e. preventing people from cocaine), it is not a prevention for the COVID-19 disease itself.

Examples:

Tweet	Justification
"Bleach sleeping pads and masks: What the military and Veterans Affairs are asking for to combat coronavirus... https://t.co/ljmyUkWF0E."	This tweet shall be marked as a true prevention as it links out to a credible news source that implicitly mentions preventative guidelines by the CDC to assist in stopping the spread of the virus.
"Personally my daily life really hasn't changed any. I believed in good hygiene before coronavirus and I will after. I already had an appropriate stock of hand sanitizer, antibacterial wipes, Lysol spray, bleach, body washes, and house hold cleaners. Its called hygiene."	This tweet shall be marked as a true prevention as it encourages good hygiene which is endorsed by WHO.

4. Fake Cure

A tweet shall be classified as a fake cure if the content *endorses a cure* and any of the following conditions are met:

1. The cure **cannot** be *verified by the World Health Organization (WHO) site*.
2. The cure **cannot** be *verified by the Center of Disease Control and Prevention (CDC)* site.
3. The cure is **not** *supported by a peer-reviewed scientific journal* that appears in Ulrich's Global Serials Directory as both "Active" and "Refereed/Peer-reviewed".
4. The cure is **not** *supported by a publicly posted working paper* authored or co-authored by tenure track faculty at a university in the top-800 universities worldwide according to the Times Higher Education World University Rankings 2019.

Examples:

Tweet	Justification
"Leaked medical conference documents reveal US hospitals preparing for 96 million coronavirus infections and 480,000 deaths! ARE YOU PREPARED? PREPARE, PREVENT, CURE NOW: 1. With Colloidal Silver! https://keto-longevity.com/colloidal-silver-for-longevity/. 2. 7,000 Vitamin D daily. 3. Get Masks+Goggles"	This tweet shall be marked as a fake cure as it is advertising a cure not endorsed/verified by WHO or CDC.

5. Fake Treatment

A treatment is different from cure as treatment improves a condition rather than completely remove the disease. A tweet shall be classified as a fake treatment if the content *endorses a treatment* and any of the following conditions are met:

1. The treatment **cannot** be *verified by the World Health Organization (WHO) site*.
2. The treatment **cannot** be *verified by the Center of Disease Control and Prevention (CDC)* site.
3. The treatment is **not** *supported by a peer-reviewed scientific journal* that appears in Ulrich's Global Serials Directory as both "Active" and "Refereed/Peer-reviewed".
4. The treatment is **not** *supported by a publicly posted working paper* authored or co-authored by tenure track faculty at a university in the top-800 universities worldwide according to the Times Higher Education World University Rankings 2019.

Examples:

Tweet	**Justification**
"I currently have the flu (haven't been tested for covid 19) and although I'm not saying that essential oils cure or protect from it eucalyptus and tea tree oil are sure helping me reduce a lot of my symptoms. When a vaccine comes I'm gonna gobble that shit up though https://t.co/PpxeRRtiue"	This tweet shall be marked as a fake treatment as it is suggesting a treatment not endorsed by WHO or CDC.
"How to get rid of Uneasiness in Breathing a symptom of Covid-19. Please use Sesame oil , Rock Salt as mentioned in Charaka Samhitha. Ayurveda has many solutions Source: Charaka Samhitha available Online"	This tweet shall be marked as a fake treatment as it is advertising a treatment not endorsed by WHO or CDC.

6. False Fact or Prevention

A tweet shall be classified as a false fact or prevention if the content mentions a false fact related to "killing" or "disrupting" coronavirus. A tweet shall be classified as a false fact or prevention if the content implicitly or explicitly *endorses a method of prevention for coronavirus* and any of the following conditions are met:

1. The prevention **cannot** be *verified by the World Health Organization (WHO) site*.
2. The prevention **cannot** be *verified by the Center of Disease Control and Prevention (CDC)* site.
3. The prevention is **not** *supported by a peer-reviewed scientific journal* that appears in Ulrich's Global Serials Directory as both "Active" and "Refereed/Peer-reviewed".

4. The prevention is **not** *supported by a publicly posted working paper* authored or co-authored by tenure track faculty at a university in the top-800 universities worldwide according to the Times Higher Education World University Rankings 2019.

Examples:

Tweet	Justification
"I heard the best way to prevent coronavirus is to pour bleach directly into your eyes and drink a full bottle of hand sanitizer."	Technically this tweet is both *sarcasm* and *fake prevention*. For the purposes of this project, this shall be coded under false fact or prevention because: i) there are no obvious signs of sarcasm such as an emoticon :) ; and (ii) disinformation is often spread as anecdotes.
"2008 Research paper demonstrating various essential oil effectivity in disrupting SARS-CoV and HSV-1 replication. #coronavirus #COVID-19 #COVID19 #CoronaVirus2020 #HSV #essentialoil #essentialoils #sars #sarscov #sars_cov #Covidcure #Hydroxychloroquine https://t.co/tke32spM8E https://t.co/drzBn2nVGp"	This tweet will be marked as false fact or prevention as it is directly against the WHO guidance, endorses a common misinformation related to essential oils, and it tries to get credibility by listing a link to a related publication.
"Garlic may help? #Covid19 Antimicrobial properties https://t.co/Anfc5SvfEy"	This tweet will be marked as false fact or prevention as it is directly against the WHO guidance, and it tries to get credibility by incorrectly listing their URL.
"@CNN These things are to 'reinforce and boost' inmune system. Ginger, Onions, Garlic: anti-bacterial, anti-fungal and anti-viral properties, is known to reduce inflammation in the body. How can you tell there is no evidence? #coronavirus"	This tweet shall be marked as a false fact or prevention as it is endorsing a prevention (via boosting of immune system) which is not supported by WHO or CDC or any scientific study.

7. Correction/Calling out

A tweet shall be classified as correction if any of the following conditions are met:

1. The tweet *calls out or makes fun of* a fake cure, a fake prevention, fake treatment, or a conspiracy theory.
2. The tweet *links out to a site that debunks, calls out or makes fun of* a fake cure, a fake prevention, fake treatment, or a conspiracy theory.
3. The tweet *calls out or make fun of violations* of social distancing rules or public health responses.

4. The tweet reports/quotes a (news) story related to consequences of a false fact, fake prevention, fake cure, fake treatment, or conspiracy theory.
5. The tweet reports/quotes a (news) story debunking a false fact, fake prevention, fake cure, fake treatment, or conspiracy theory.

Examples:

Tweet	**Justification**
"Taking a hot bath, eating lots of garlic and spraying chlorine all over your body - these are just some of the so-called solutions to Coronavirus that can be found on the internet. https://t.co/CKQhmAKPRX"	This tweet can be classified as correction/calling out as it points to a site titled: *Coronavirus: So-called 'solutions' debunked by World Health Organisation*. The usage of the term *"so-called"* in the tweet also indicates that this is a correction.
"No, #Cocaine does not protect against #coronavirus -French Officials."	This tweet can be classified as a correction/calling out as it endorses the statement by french officials calling out a specific fake prevention.
"Someone told me, in support of a conspiracy theory, "well on the side of a bottle of bleach it says it kills coronavirus" and I'm like...y'all there are lots of coronaviruses that is not how this works."	This tweet can be classified as a correction/calling out as the author clearly describes their stance by calling out the fake cure/prevention related to drinking bleach.
"Coronavirus myths, debunked: A cattle vaccine, bioweapons and a $3,000 test https://t.co/ykoLUlGspQ"	This tweet should be marked as a correction/calling out since it links out to a credible news source debunking the claims.
"Another day another meme to debunk: Vaccines for the bovine coronavirus will not cure COVID-19 https://t.co/qwHkLOnXw4"	This tweet should be marked as a correction/calling out since it links out to a credible news source debunking the claims.
"This is what 'the cure can't be worse than disease' crowd is ok with happening. https://t.co/0YSv1C69St."	This tweet should be marked as a correction/calling out since it calls out a public health response and social distancing, linking to a credible news source.
"Coronavirus: Cocaine cure myth spreads, rebutted by French government - Business Insider https://t.co/rzusol40YG"	This tweet should be marked as a correction/calling out since it endorses a public health response by the French government to debunk misinformation.

8. Sarcasm/Satire

A tweet shall be classified as sarcasm/satire if any of the following conditions are met

1. The tweet contains *clear signs of a satire* calling out a fake cure, a fake prevention or a

conspiracy.

2. The tweet includes a clear *joke* about a fake cure, a fake prevention or a conspiracy.

Concretely, this is a tweet where the information in the post is false but is presented using humor, irony, exaggeration, or ridicule to expose and criticize people's stupidity or vices, particularly in the context of contemporary politics and other topical issues. This kind of post is used to ridicule other false statements or people.

Examples:

Tweet	**Justification**
"Which essential oil is best for getting people to relax about the Coronavirus?"	This tweet can be classified as a satire as it is using sarcasm to call out the essential oil fake cures.
"Sesame oil, oregano oil and garlic. Who needs vaccines when you can marinate?"	This tweet can be classified as a satire as it is using sarcasm to call out the different fake cures.
"There's no way I can get the Coronavirus I snorted cocaine off the back of the toilet at alahome I'm immune to death itself."	This tweet can be classified as a satire as it is using sarcasm to call out the cocaine-related fake cure/treatment.
"Elmo isn't scared of the Coronavirus. Elmo's theory is that if you do enough cocaine, it'll kill the virus. https://t.co/YX6gsYAVI0"	This tweet can be classified as a satire as it is using sesame street character Elmo, and sarcasm to call out the cocaine related fake cure/treatment. If you follow the link in the tweet, it also shows a clearly funny image of Elmo snorting cocaine signalling that this tweet should be marked as sarcasm/satire.

9. True Public Health Response

A tweet shall be classified as true public health response if it is a statement about the public health response (eg. changes to essential services, location of testing, pending lockdown, etc.), and it *links to a mainstream news source or official government website* through which the statement can be verified.

Examples:

Tweet	Justification
"'Ontario and Quebec designate alcohol producers and retailers as essential services during current COVID-19 crisis... https://t.co/EjoBZMDUy2'"	This tweet should be marked as a true public health response as it links out to the official Quebec government website.
"Centers for Disease Control (CDC) and Prevention webpage for CORONAVIRUS (COVID-19) https://t.co/6QueBAGPYL #KYSPIN... https://t.co/M4ovNeWhNz"	This tweet should be marked as a true public health response as it links out to the CDC website on Coronavirus.
"Kudos @Unilever 'Free soap, sanitiser, bleach and food to the value of 100 million' '500 million of cash flow relief to support livelihoods' 'We will cover our employees, contractors and others who we manage or who work on our sites' https://t.co/cwPsNlVoXF #coronavirus https://t.co/sZHUR1uHU4"	This tweet should be marked as a true public health response as the statement comes from Unilever's public health response for its customers, and stakeholders.

10. False Public Health Response

A tweet shall be classified as false public health response if it makes a claim about the public health response (eg. changes to essential services, location of testing, pending lockdown, etc.), but the claim *cannot be verified by a credible news source*.

Examples:

Tweet	Justification
"Russia presents Covid-19 TREATMENT based on anti-malaria drug https://t.co/D8mTTm6g3n"	This tweet should be marked as a false public health response as it cannot be verified by the link provided that directs to a suspicious Russian news site.
"Italy allows malaria and HIV drugs for coronavirus treatment https://t.co/dqe91DgPv5 https://t.co/EoeURlleUX"	This tweet should be marked as a false public health response as it cannot be verified by the link provided that directs to a suspicious Russian news site.

11. Politics

A tweet shall be classified as politics if the tweet *mentions a political individual, institution, or government organization (eg. Congress, Democratic or Republican party)*, and any of the

following conditions are met:

1. The tweet *implicitly comments* on actions taken by the political actor.
2. The tweet *provides commentary* on actions taken by the political actor.

Examples:

Tweet	Justification
"'Trump kept saying it was basically pretty much a cure': Woman whose husband died after ingesting chloroquine warns the public not to 'believe anything that the president says' https://t.co/hWo6Zc4aOw"	This tweet should be marked as politics, since it mentions a political actor (Trump) and is implicitly commenting on that actor's statements.
"Biden on Coronavirus: 'We Have to Take Care of the Cure – That Will Make the Problem Worse No Matter What' https://t.co/Sn0CxREPJU"	This tweet should be marked as politics, since it mentions a political actor (Biden) and provides commentary on his action.
"Is Nancy Pelosi serious? Covid-19 bill has $300 million for Sesame Street & National Endowment of Arts? An untold... https://t.co/AUNx3SIEkl"	This tweet should be marked as politics, since it mentions a political actor (Nancy Pelosi) and responds to an action taken by her (incorporated funding for these groups in the COVID-19 bill).
"@realDonaldTrump @nytimes #Trump defunded & eliminated the #Pandemic Research &; Prevention Department. #Coronavirus... https://t.co/F5PZ6quYd1" *"Nailed it! ? God why aren't Republicans praying for a cure for .the #Coronavirus rather than sacrificing grandpare... https://t.co/njYBhofYYi."*	This tweet should be marked as politics, since it mentions a political actor (Trump) and responds to an action taken by him (defunded Pandemic research). This tweet should be marked as politics, since it mentions a political party (Republican) and its supporters in relation to COVID-19.
"@Brent68189672 @rodgregg66 @theOriginalOWL @KurisuS @RudyGiuliani Trump never said to take fish tank cleaner to remove your COVID infection. He said the drug for malaria treatment offers hope. If you think I'm stupid you must also think that Cuomo is stupid for trialing it on New Yorkers. https://t.co/6wNZiTOBM4"	This tweet can be both politics or fake treatment, but it should be marked as politics, since the main theme of the tweet is Trump's statement which is what is being disputed rather than the malaria drug itself.

12. Ambiguous/Difficult to Classify

A tweet shall be classified as ambiguous if the stance of the author is not clear, and the post can potentially fall into either of the contrasting categories (eg. true treatment vs. false treatment, or true prevention vs. false prevention).

Examples:

Tweet	Justification
"Is it true that the tropics insulates us from covid 19? Is it true garlic is the magic bullet? Who are most vulnerable who are less? What best first aids? What are the numbers of first responders to call? Where do you go to report suspected cases? All these are not clear"	This tweet shall be classified as ambiguous as the author is asking a question on the validity of different fake treatments and fake preventions, and the stance of the author is not clear.
"@nevadazmom @Meadmommy @RudyGiuliani Lady our entire country is shut down. There are very good medical reasons from Medical professionals to believe that hydroxychloroquine or Z-pack combo could save thousands of lives and end a looming depression that could kill millions. Why isn't this the most important thing?"	This tweet shall be classified as ambiguous as the stance of the author is not clear.

13. Commercial Activity or Promotion

A tweet shall be classified as commercial activity or promotion if it includes a company advertising or selling coronavirus-related protective and preventative gear (eg. hand sanitizers, face masks, cleaners).

Examples:

Tweet	Justification
"Fight Coronavirus, disinfect your home with hypo bleach...#HypoFightCoronaVirus .#CoronaVirusUpdate .#hypoGoWipeo https://t.co/lMvCrQ9KIz"	This tweet shall be classified as commercial activity or promotion as it comes from a cleaning supplies company and is promoting the purchase of their product.
"ARE YOU IN NEED OF COVID-19 CORONAVIRUS CDC..SUPPLIES ? WE SELL IN BULK SANITIZERS MASKS GLOVES..N95 BLEACH TISSUE... https://t.co/k9GAI1iD5D"	This tweet shall be classified as commercial activity or promotion as it comes from a (probable) bot account linking to a suspicious selling website.

14. Emergency Response

A tweet shall be classified in this category if it mentions a viable emergency response (eg. changes in government funding for education programs, links to mental health resource).

Examples:

Tweet	Justification
"These days can be difficult.. ?? asking for help is brave. ..Suicide Prevention.800-273-8255..Substance Abuse/Menta... https://t.co/OiQPPjQBLL."	This tweet should be coded under this category as it links out to advice on mental health during the pandemic.

15. News

A tweet shall be classified as news if it cannot be classified in any of the other categories, and it quotes a news story and links to a news site.

Note: If the theme of the news story is about any of the above categories, it should be classified under that category. For example, if the news story is about debunking myths related to COVID-19 or a person dying of a fake cure, that should be classified as Correction/Calling Out. Similarly, if the news story is about panic buying, it should be classified as panic buying.

Examples:

Tweet	Justification
"'Modern planning and civil engineering were born out of the mid-19th century development of sanitation in response to the spread of malaria and cholera in cities. Digital infrastructure might be the sanitation of our time.' https://t.co/561HN98RrP"	This tweet should be coded under news category as it talks about the role of urban planning on handling pandemics, and directly quotes the CityLab's article on this topic.
"10 new Utah #coronavirus cases reported tonight—nine in Salt Lake County—bringing our total to 19. Stay home, people! https://t.co/Nvn4FidE0E"	This tweet should be coded under news category as it quotes the Utah government site reporting the number of coronavirus cases in Salt Lake County.

16. Panic Buying

A tweet shall be classified in this category if it mentions or comments on panic buying or its consequences in the context of COVID-19. A tweet shall also be classified in this category if it quotes a news site/story that talks about panic buying.

Examples:

Tweet	Justification
"Found two bottles of bleach at @Target yesterday. I've never been happier then finding bottles of bleach before. #thelittlethings #COVID-19 #coronavirus"	This tweet should be coded under this category as it comments on the shortage of bleach caused by panic buying
"No milk, no bleach: Americans awake to coronavirus panic buying https://t.co/LKDBkZilVg"	This tweet should be coded under this category as it comments on the shortage of bleach caused by panic buying
"Toilet paper, the surprise coin of the realm of the coronavirus outbreak, was gone from aisle 3. Most laundry detergent, bleach and cleanser were gone from aisle 5. https://t.co/X3znsymeEb"	This tweet should be coded under this category as it comments on the shortage of bleach caused by panic buying

A.3 Additional Notes

1. If it is not clear whether a certain object is mentioned within the context of a treatment, prevention or cure, but is essentially false, it should be classified as a false fact or prevention. For example, essential oils have been mentioned as a "prevention" as well as a "cure" to COVID-19. If the tweet does not explicitly mention if it is a cure or prevention, it should be classified as a "false fact or prevention".
2. If a tweet falls into more than one category, try to infer the theme of the tweet. For example, the tweet: *"Trump never said to take fish tank cleaner to remove your COVID infection. He said the drug for malaria treatment offers hope. If you think I'm stupid you must also think that Cuomo is stupid for trialing it on New Yorkers."* can be classified as fake treatment or politics, but it should be marked as politics, since the main theme of the tweet is Trump's statement which is what is being disputed rather than the malaria drug itself.

Bibliography

[1] Public health echo chambers in a time of mistrust & misinformation - digital health @ harvard, february 2017. URL https://cyber.harvard.edu/events/digitalhealth/2017/02/GyenesSeymour. 1, 2.4.2

[2] Cook county officials report measles exposure in city, suburbs. url=https://news.wttw.com/2018/07/23/cook-county-officials-report-measles-exposure-city-suburbs. 1

[3] Morbidity and mortality weekly report (mmwr). url=https://tinyurl.com/cdc-morbidity, Nov 2018. 1

[4] Measles (rubeola). url=https://www.cdc.gov/measles/cases-outbreaks.html, Nov 2018. 1

[5] Muhammad Abdul-Mageed, AbdelRahim Elmadany, Dinesh Pabbi, Kunal Verma, and Rannie Lin. Mega-cov: A billion-scale dataset of 65 languages for covid-19. *arXiv preprint arXiv:2005.06012*, 2020. 3.1.1

[6] Firoj Alam, Shaden Shaar, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Kareem Darwish, et al. Fighting the covid-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. *arXiv preprint arXiv:2005.00033*, 2020. 3.1.1

[7] Joachim Allgaier and Anna Lydia Svalastog. The communication aspects of the ebola virus disease outbreak in western africa–do we need to counter one, two, or many epidemics? *Croatian medical journal*, 56(5):496, 2015. 1

[8] Sarah Alqurashi, Ahmad Alhindi, and Eisa Alanazi. Large arabic twitter dataset on covid-19. *arXiv preprint arXiv:2004.04315*, 2020. 3.1.1

[9] Neal Altman, Kathleen M Carley, and Jeffrey Reminga. Ora user's guide 2017. *Carnegie-Mellon Univ. Pittsburgh PA Inst of Software Research International, Tech. Rep.*, 2017. 3.4.2

[10] Neal Altman, Kathleen M Carley, and Jeffrey Reminga. Ora user's guide 2018. *Carnegie-Mellon Univ. Pittsburgh PA Inst of Software Research International, Tech. Rep.*, 2018. 2.1, 2.4.2, 3.4.2, 3.2, 3.3

[11] Neal Altman, Kathleen M Carley, and Jeffrey Reminga. Ora user's guide 2020. *Carnegie-Mellon Univ. Pittsburgh PA Inst of Software Research International, Tech. Rep.*, 2020. 2.3.2

[12] Oxford Analytica. Misinformation will undermine coronavirus responses. *Emerald Expert Briefings*, (oxan-db), 2020. 1

[13] Amit Arora and Robin Wendell Evans. Dental caries in children: a comparison of one non-fluoridated and two fluoridated communities in nsw. *New South Wales public health bulletin*, 21(12):257–262, 2011. 1

[14] Darrin Baines, RJ Elliott, et al. Defining misinformation, disinformation and malinformation: An urgent need for clarity during the covid-19 infodemic. *Discussion Papers*, pages 20–06, 2020. 1

[15] Juan M Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and Gerardo Chowell. A large-scale covid-19 twitter chatter dataset for open scientific research–an international collaboration. *arXiv preprint arXiv:2004.03688*, 2020. 3.1.1

[16] David Beskow and Kathleen M Carley. *Social Cybersecurity*. Springer, 2020. 3.4.3

[17] David Beskow, Kathleen M Carley, Halil Bisgin, Ayaz Hyder, Chris Dancy, and Robert Thomson. Introducing bothunter: A tiered approach to detection and characterizing automated activity on twitter. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Springer*, 2018. 3.4.3

[18] David M Beskow and Kathleen M Carley. Bot-hunter: a tiered approach to detecting & characterizing automated activity on twitter. In *Conference paper. SBP-BRiMS: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, 2018. 3.4.3

[19] David M Beskow and Kathleen M Carley. Bot conversations are different: leveraging network metrics for bot detection in twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 825–832. IEEE, 2018. 3.4.3

[20] Beth L Boser. Mothers' anti-vax to pro-vax conversions. *Recovering Argument*, page 21, 2018. 2.3.1, 4

[21] James J Bradac, Anthony Mulac, and Sandra A Thompson. Men's and women's use of intensifiers and hedges in problem-solving interaction: Molar and molecular analyses. *Research on Language and Social Interaction*, 28(2):93–116, 1995. 2.4.1

[22] J Scott Brennen, Felix Simon, Philip N Howard, and Rasmus Kleis Nielsen. Types, sources, and claims of covid-19 misinformation. *Reuters Institute*, 7, 2020. 3.1.2

[23] Petter Brodin. Why is covid-19 so mild in children? *Acta Paediatrica*, 109(6):1082–1083, 2020. 4

[24] David A Broniatowski, Karen M Hilyard, and Mark Dredze. Effective vaccine communication during the disneyland measles outbreak. *Vaccine*, 34(28):3225–3228, 2016. 2.1, 2.2.2

[25] David A Broniatowski, Amelia M Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C Quinn, and Mark Dredze. Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American journal of public health*, 108 (10):1378–1384, 2018. 2.2.1, 2.2.2, 3.4.5

[26] David A Broniatowski, Daniel Kerchner, Fouzia Farooq, Xiaolei Huang, Amelia M Jami-

son, Mark Dredze, and Sandra Crouse Quinn. The covid-19 social media infodemic reflects uncertainty and state-sponsored propaganda. *arXiv preprint arXiv:2007.09682*, 2020. 3.1.2

[27] Stephen J Brown, William Goetzmann, Roger G Ibbotson, and Stephen A Ross. Survivorship bias in performance studies. *The Review of Financial Studies*, 5(4):553–580, 1992. 2.2

[28] John Bryden, Sebastian Funk, and Vincent AA Jansen. Word usage mirrors community structure in the online social network twitter. *EPJ Data Science*, 2(1):3, 2013. 2.1

[29] Kathleen M Carley. Ora: A toolkit for dynamic network analysis and visualization., 2017. 2.1, 2.4.2, 3.2, 3.3

[30] L Richard Carley, Jeff Reminga, and Kathleen M Carley. Ora & netmapper. 3.4.2

[31] Timothy Caulfield, Alessandro R Marcon, Blake Murdoch, Jasmine M Brown, Sarah Tinker Perrault, Jonathan Jarry, Jeremy Snyder, Samantha J Anthony, Stephanie Brooks, Zubin Master, et al. Health misinformation and the power of narrative messaging in the public sphere. *Canadian Journal of Bioethics/Revue canadienne de bioéthique*, 2(2):52–60, 2019. 4

[32] Man-pui Sally Chan, Christopher R Jones, Kathleen Hall Jamieson, and Dolores Albarracín. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science*, 28(11):1531–1546, 2017. 1

[33] 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. 3.1.1

[34] Wen-Ying Sylvia Chou, April Oh, and William MP Klein. Addressing health-related misinformation on social media. *JAMA*, 2018. 1

[35] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The covid-19 social media infodemic. *arXiv preprint arXiv:2003.05004*, 2020. 1

[36] Limeng Cui and Dongwon Lee. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*, 2020. 3.1.1

[37] Enyan Dai, Yiwei Sun, and Suhang Wang. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 853–862, 2020. 3.1.1

[38] Sonya Dal Cin, Mark P Zanna, and Geoffrey T Fong. Narrative persuasion and overcoming resistance. *Resistance and persuasion*, 2:175–191, 2004. 4

[39] Arkin R Dharawat, Ismini Lourentzou, Alex Morales, and ChengXiang Zhai. Drink bleach or do what now? covid-hera: A dataset for risk-informed health decision making in the presence of covid19 misinformation. 2020. 3.1.1

[40] Mark Dredze, David A Broniatowski, and Karen M Hilyard. Zika vaccine misconceptions: A social media analysis. *Vaccine*, 34(30):3441, 2016. 2.1

[41] Mark Dredze, David A Broniatowski, Michael C Smith, and Karen M Hilyard. Under-

standing vaccine refusal: why we need social media now. *American journal of preventive medicine*, 50(4):550–552, 2016. 2.1

[42] Mark Dredze, Zachary Wood-Doughty, Sandra Crouse Quinn, and David A Broniatowski. Vaccine opponents' use of twitter during the 2016 us presidential election: Implications for practice and policy. *Vaccine*, 35(36):4670–4672, 2017. 2.2.1, 2.2.2

[43] Nikita Duseja and Harsh Jhamtani. A sociolinguistic study of online echo chambers on twitter. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 78–83, 2019. 2.1, 2.3.1, 2.4.1, 3.4.4

[44] Ash Evans. Stance and identity in twitter hashtags. *Language@ internet*, 13(1), 2016. 2.2.2

[45] Emilio Ferrara. What types of covid-19 conspiracies are populated by twitter bots? *First Monday*, 2020. 3.1.2

[46] Alison Gibbons and Andrea Macrae. *Pronouns in literature: Positions and perspectives in language*. Springer, 2018. 2.3.1

[47] Katherine Giuffre. Cultural production in networks. 2015. 2.3.2

[48] Arthur C Graesser, Danielle S McNamara, Zhiqang Cai, Mark Conley, Haiying Li, and James Pennebaker. Coh-metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, 115(2):210–229, 2014. 3.4.4

[49] Sedigheh Khademi Habibabadi and Pari Delir Haghighi. Topic modelling for identification of vaccine reactions in twitter. In *Proceedings of the Australasian Computer Science Week Multiconference*, page 31. ACM, 2019. 2.1

[50] E Allison Hagood and Stacy Mintzer Herlihy. Addressing heterogeneous parental concerns about vaccination with a multiple-source model: a parent and educator perspective. *Human vaccines & immunotherapeutics*, 9(8):1790–1794, 2013. 2.1, 2.4.2, 2.5

[51] Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. Arcov-19: The first arabic covid-19 twitter dataset with propagation networks. *arXiv preprint arXiv:2004.05861*, 2020. 3.1.1

[52] Molly Harbarger. Third person with measles found in multnomah county – latest brings outbreak to 4. url=https://www.oregonlive.com/health/ index.ssf/2018/07/third_person_with_measles_foun.html, Jul 2018. 1

[53] Beth L Hoffman, Elizabeth M Felter, Kar-Hai Chu, Ariel Shensa, Chad Hermann, Todd Wolynn, Daria Williams, and Brian A Primack. It's not all about autism: The emerging landscape of anti-vaccination sentiment on facebook. *Vaccine*, 37(16):2216–2223, 2019. 2.1

[54] Lawrence A Hosman. The evaluative consequences of hedges, hesitations, and intensifies: Powerful and powerless speech styles. *Human communication research*, 15(3):383–406, 1989. 2.4.1

[55] Binxuan Huang and Kathleen M Carley. Disinformation and misinformation on twitter during the novel coronavirus outbreak. *arXiv preprint arXiv:2006.04278*, 2020. 3.1.2

[56] Xiaolei Huang, Amelia Jamison, David Broniatowski, Sandra Quinn, and Mark Dredze.

Coronavirus twitter data: A collection of covid-19 tweets with automated annotations, 2020. 3.1.1

[57] Daniel Jolley and Karen M Douglas. The effects of anti-vaccine conspiracy theories on vaccination intentions. *PloS one*, 9(2):e89177, 2014. 2.1

[58] Anna Kata. A postmodern pandora's box: anti-vaccination misinformation on the internet. *Vaccine*, 28(7):1709–1716, 2010. 2.1

[59] Suin Kim, Ingmar Weber, Li Wei, and Alice Oh. Sociolinguistic analysis of twitter in multilingual societies. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 243–248, 2014. 2.1

[60] David Krackhardt and Robert N Stern. Informal networks and organizational crises: An experimental simulation. *Social psychology quarterly*, pages 123–140, 1988. 2.3.2

[61] Siddique Latif, Muhammad Usman, Sanaullah Manzoor, Waleed Iqbal, Junaid Qadir, Gareth Tyson, Ignacio Castro, Adeel Razi, Maged N Kamel Boulos, Adrian Weller, et al. Leveraging data science to combat covid-19: A comprehensive review. 2020. 3.1.1

[62] Gabrielle Levy. Public confidence in vaccines sags, new report finds. url=https://www.usnews.com/news/health-care-news/articles/2018-05-21/public-confidence-in-vaccines-sags-new-report-finds. 1

[63] Noni E MacDonald, Robb Butler, and Eve Dubé. Addressing barriers to vaccine acceptance: an overview. *Human vaccines & immunotherapeutics*, 14(1):218–224, 2018. 2.1

[64] Shahan Ali Memon and Kathleen M. Carley. Cmu-miscov19: A novel twitter dataset for characterizing covid-19 misinformation, Sep 2020. 3.3

[65] Shahan Ali Memon and Kathleen M Carley. Characterizing covid-19 misinformation communities using a novel twitter dataset. *arXiv preprint arXiv:2008.00791*, 2020. URL https://arxiv.org/abs/2008.00791. 1

[66] Shahan Ali Memon, Aman Tyagi, David R Mortensen, and Kathleen M Carley. Characterizing sociolinguistic variation in the competing vaccination communities. *arXiv preprint arXiv:2006.04334*, 2020. URL https://arxiv.org/abs/2006.04334. 1, 2.1, 3.4.2, 3.4.4, 3.6

[67] Tanushree Mitra, Scott Counts, and James W Pennebaker. Understanding anti-vaccination attitudes in social media. In *Tenth International AAAI Conference on Web and Social Media*, 2016. 2.1

[68] Kevin Munger. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3):629–649, 2017. 4

[69] John Nerbonne. The secret life of pronouns. what our words say about us. *Literary and Linguistic Computing*, 29(1):139–142, 2014. 2.4.1

[70] Catherine Ordun, Sanjay Purushotham, and Edward Raff. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. *arXiv preprint arXiv:2005.03082*, 2020. 3.1.2

[71] James W Pennebaker. The secret life of pronouns. *New Scientist*, 211(2828):42–45, 2011.

3.4.4

[72] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015. 3.4.4, 3.4.4

[73] Umair Qazi, Muhammad Imran, and Ferda Ofli. Geocov19: a dataset of hundreds of millions of multilingual covid-19 tweets with location information. *SIGSPATIAL Special*, 12 (1):6–15, 2020. 3.1.1

[74] Joe Raedle. Health alert issued after child with measles visited boston - the boston globe. url=https://www.bostonglobe.com/metro/2018/07/27/health-alert-issued-after-child-with-measles-visited-boston/IBRFn8EUnFFdAQfrJJOewK/y.html, Jul 2018. 1

[75] Angela Giuffrida in Rome. Italy's five star movement blamed for surge in measles cases. url=https://www.theguardian.com/world/2017/mar/23/italys-five-star-movement-blamed-for-surge-in-measles-cases, Mar 2017. 1

[76] Angeline Sangalang, Yotam Ophir, and Joseph N Cappella. The potential for narrative correctives to combat misinformation. *Journal of communication*, 69(3):298–319, 2019. 3.6, 4

[77] Gautam Kishore Shahi and Durgesh Nandini. Fakecovid–a multilingual cross-domain fact check news dataset for covid-19. *arXiv preprint arXiv:2006.11343*, 2020. 3.1.1

[78] Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. An exploratory study of covid-19 misinformation on twitter. *arXiv preprint arXiv:2005.05710*, 2020. 3.1.2

[79] Karishma Sharma, Sungyong Seo, Chuizheng Meng, Sirisha Rambhatla, and Yan Liu. Covid-19 on social media: Analyzing misinformation in twitter conversations. *arXiv preprint arXiv:2003.12309*, 2020. 3.1.2

[80] Ashley Shelby and Karen Ernst. Story and science: how providers and parents can utilize storytelling to combat anti-vaccine misinformation. *Human vaccines & immunotherapeutics*, 9(8):1795–1801, 2013. 2.1, 4

[81] Junaid Shuja, Eisa Alanazi, Waleed Alasmary, and Abdulaziz Alashaikh. Covid-19 open source data sets: A comprehensive survey. *medRxiv*, 2020. 3.1.1

[82] Neil J Smelser, Paul B Baltes, et al. *International encyclopedia of the social & behavioral sciences*, volume 11. Elsevier Amsterdam, 2001. 2.3.2

[83] Tara C Smith. Vaccine rejection and hesitancy: a review and call to action. In *Open forum infectious diseases*, volume 4. Oxford University Press, 2017. 2.1, 2.4.2, 2.5

[84] Xingyi Song, Johann Petrak, Ye Jiang, Iknoor Singh, Diana Maynard, and Kalina Bontcheva. Classification aware neural topic model and its application on a new covid-19 disinformation corpus. *arXiv preprint arXiv:2006.03354*, 2020. 3.1.1

[85] Andy SL Tan, Chul-joo Lee, and Jiyoung Chae. Exposure to health (mis) information: Lagged effects on young adults' health behaviors and potential pathways. *Journal of Communication*, 65(4):674–698, 2015. 1

[86] Aman Tyagi, Mathew Babcock, Kathleen M. Carley, and Douglas C. Sicker. Polarizing tweets on climate change. *To appear in International Conference SBP-BRiMS*, 2020. 2.2.2

[87] Aman Tyagi, Anjalie Field, Priyank Lathwal, Yulia Tsvetkov, and Kathleen M. Carley. A computational analysis of polarization on indian and pakistani social media, 2020. 2.2.2

[88] Marcela Veselková. Narrative policy framework: Narratives as heuristics in the policy process. *Human Affairs*, 27(2):178, 2017. 3.4.4

[89] Ramon Villa-Cox, Sumeet Kumar, Matthew Babcock, and Kathleen M Carley. Stance in replies and quotes (srq): A new dataset for learning stance in twitter conversations. *arXiv preprint arXiv:2006.00691*, 2020. 3.1.1

[90] Lindsey Wahowiak. Public health working to fight misinformation through trust, relationships: Facts not enough. url=http://thenationshealth.aphapublications.org/content/48/5/1.2, Jul 2018. 1

[91] Stanley Wasserman, Katherine Faust, et al. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994. 2.3.2

[92] Zhu Xiaojin and Ghahramani Zoubin. Learning from labeled and unlabeled data with label propagation. *Tech. Rep., Technical Report CMU-CALD-02–107, Carnegie Mellon University*, 2002. 2.2.2

[93] Kai-Cheng Yang, Christopher Torres-Lugo, and Filippo Menczer. Prevalence of low-credibility information on twitter during the covid-19 outbreak. *arXiv preprint arXiv:2004.14484*, 2020. 3.1.2

[94] Lynne Young and Claire Harrison. *Systemic functional linguistics and critical discourse analysis: Studies in social change*. A&C Black, 2004. 2.4.1

[95] Xiaoyi Yuan, Ross J Schuchard, and Andrew T Crooks. Examining emergent communities and social bots within the polarized online vaccination debate in twitter. *Social Media+ Society*, 5(3):2056305119865465, 2019. 2.1

[96] Engku Nuraishah Huda E Zainudin, Khairool Azizul Mohammad, Athirah Aris, Intan Azura Shahdan, and Pahang Kuantan. Vaccination: Influencing factors and view from an islamic perspective. 2.1

[97] Zhen Zhao and Elizabeth T Luman. Progress toward eliminating disparities in vaccination coverage among us children, 2000–2008. *American journal of preventive medicine*, 38(2): 127–137, 2010. 2.1

[98] Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. Recovery: A multimodal repository for covid-19 news credibility research. *arXiv preprint arXiv:2006.05557*, 2020. 3.1.1