AAAS
Health Data Science
Volume 2021, Article ID 9837856, 11 pages
https://doi.org/10.34133/2021/9837856

Health Data Science
A SCIENCE PARTNER JOURNAL

Research Article

Characterizing Discourse about COVID-19 Vaccines: A Reddit Version of the Pandemic Story

Wei Wu,[1] **Hanjia Lyu**,[1] **and Jiebo Luo**[2]

[1]*Goergen Institute for Data Science, University of Rochester, Rochester, USA*
[2]*Department of Computer Science, University of Rochester, Rochester, USA*

Correspondence should be addressed to Jiebo Luo; jluo@cs.rochester.edu

Received 6 April 2021; Accepted 28 June 2021; Published 27 August 2021

It has been one year since the outbreak of the COVID-19 pandemic. The good news is that vaccines developed by several manufacturers are being actively distributed worldwide. However, as more and more vaccines become available to the public, various concerns related to vaccines become the primary barriers that may hinder the public from getting vaccinated. Considering the complexities of these concerns and their potential hazards, this study is aimed at offering a clear understanding about different population groups' underlying concerns when they talk about COVID-19 vaccines—particularly those active on Reddit. The goal is achieved by applying LDA and LIWC to characterize the pertaining discourse with insights generated through a combination of quantitative and qualitative comparisons. Findings include the following: (1) during the pandemic, the proportion of Reddit comments predominated by conspiracy theories outweighed that of any other topics; (2) each subreddit has its own user bases, so information posted in one subreddit may not reach that from other subreddits; and (3) since users' concerns vary across time and subreddits, communication strategies must be adjusted according to specific needs. The results of this study manifest challenges as well as opportunities in the process of designing effective communication and immunization programs.

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), known as COVID-19, has hit 222 countries and caused over 2 million deaths in the worldwide as of January 14, 2021 (https://www.worldometers.info/coronavirus/). Since the World Health Organization (WHO) declared COVID-19 a pandemic (https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020), it has caused catastrophic damages to various aspects of the human society, from economy [1] to psychology [2]. To minimize the deadly impacts of the pandemic, scientists from different countries have been actively involved in the investigations of effective treatments and vaccines. Starting in January 2021, the vaccines developed by the corporations led by Pfizer and Moderna are being distributed across the U.S., while the Johnson & Johnson vaccine is also on the way (https://www.cdc.gov/coronavirus/2019-ncov/vaccines/8-things.html). Despite the arduous efforts of scientists that have demonstrated the efficacy of possible COVID-19 vaccines and their protection mechanisms to human bodies [3], the public attitudes towards COVID-19 vaccines are nothing if not disparate.

As early as June 2020, a survey covering 13,426 participants in 19 countries already showed that the potential acceptance of a COVID-19 vaccine varies largely from country to country: in China, the acceptance rate of a potentially safe and effective COVID-19 vaccine is 90%, which is the highest among all countries, while that rate in Russia is only 55% [4]. The study suggested that the variation in the vaccine acceptance rates may be a result of the difference in levels of trust in central governments. Another study [5], which mainly focused on adults in the United States, conducted an online survey with 2,006 qualified participants and investigated factors that influence their acceptability of a COVID-19 vaccine. One of the main implications of this study is that participants whose political beliefs are moderate or liberal have a relatively higher vaccine acceptance rate in this case. There are

also many other emerging studies that conducted similar surveys to analyze the acceptability of a COVID-19 vaccine among different social and ethnic groups (e.g., Goldman et al. [6]), providing policy-makers with a clearer picture of the public's attitudes towards COVID-19 vaccines. These studies are important because they shed light on the hidden obstacles and challenges in the vaccine distribution process and offer critical information for implementing appropriate communication strategies in times of crisis. Nevertheless, because of the limitations of the survey approach—including but not limited to a small sample size and a definite number of available choices—it is difficult to know what the specific concerns of the general public are when they think of COVID-19 vaccinations. In addition, out of the fear of being considered as "antivaxxers," people with concerns may choose to restrain themselves from expressing their genuine opinions in the aforementioned surveys [7] or not to participate in them at all.

To supplement and verify the findings in the preceding studies, we employ a data-mining approach with natural language processing (NLP) techniques to investigate the discourse related to COVID-19 vaccines using large-scale social media data. Since the advent of social media platforms, they have been widely used as reliable data sources for a variety of research purposes. From presidential election analyses [8] to electronic cigarette perception examinations [9], social media data demonstrates advantages such as the involvement of an extensive population and the revelation of hidden patterns that are easily overlooked otherwise. The use of social media data in topics like the public opinions towards COVID-19 vaccines has also been reported. For instance, in a recent study based on millions of Twitter data, Lyu et al. [10] found that safety, effectiveness, and political issues are chief concerns of the U.S. public when talking about COVID-19 vaccines. However, it is unclear whether the findings revealed by this study are valid with other online communities, which differ from Twitter in user bases and platform features. To better characterize the online discourse related to COVID-19 vaccines, we think it is important and necessary to extend the "ground truth" of public opinions to a broader range of online communities, such as Reddit.

Ranked as the 7th most visited website in the United States by November 2020 (https://www.alexa.com/siteinfo/reddit.com), Reddit is one of the more news-oriented social media platforms according to a research published by the Pew Research Center (https://www.journalism.org/2016/02/25/seven-in-ten-reddit-users-get-news-on-the-site/). Unlike Twitter which requests users to fill in profile information, Reddit only asks users to enter usernames in the sign-up process, which results in an anonymous environment. Because of this anonymity feature, users of Reddit, who are referred to as Redditors, perceive a freedom of expression and tend to share their true opinions on topics that they would not openly address on other social media platforms [11]. Therefore, in this work, we collect data from Reddit to examine the thematic characteristics of the discussions of COVID-19 vaccines on it. In doing so, we hope to answer two research questions:

(1) *RQ 1*: What specific topics characterize the discussions of COVID-19 vaccines on Reddit? How do the topics vary across time and subreddits?

(2) *RQ 2*: Given three most active subreddits, what can we learn from their overlapping users? How do they contribute to the thematic and linguistic similarities and differences among these subreddits?

Our approach consists of the Latent Dirichlet Allocation (LDA) topic modeling [12], the Linguistic Inquiry and Word Count (LIWC) text analysis [13], and visual comparisons. Taking into account the nature of social media data, whose development is free from the intervention of researchers, we apply the longitudinal study design to detect the changes in side-wide discussion topics on Reddit over a 9-month time period, and use the cross-sectional study design to analyze thematic, linguistic similarity, difference, and membership in three different subreddits. Both the longitudinal and cross-sectional studies are observational studies, which are widely used approaches in social science fields. Through a combination of computational and qualitative methods, we aim to generate a more comprehensive understanding of the public's concerns about COVID-19 vaccines, thus paving the way for effective communication strategies geared towards the needs of different communities.

2. Related Work

Factors that influence vaccine acceptance and coverage are complex. In the review of articles published between January 2007 and November 2012 [14], researchers examined 1,164 articles that studied vaccine hesitancy across different regions in the world, concluding that even given a definite number of determinants (i.e., level of income, education), their impacts on people's perceptions of vaccines vary across time, place, and vaccine types. The study particularly pointed out that as the access to healthcare becomes less a barrier in many countries, personal attitudes and beliefs put an increasing impact on vaccine behaviors, especially in regions wherein vaccine services are available to the most of the population. This finding coheres with what was suggested by another study, which forecasts the trends in vaccination coverage using a time-series analysis over 30 years [15]. Hence, from either a cross-sectional or a longitudinal perspective, probing into the role of personal opinions in vaccination behaviors at different settings becomes a necessity in the pursuit of a deeper understanding about factors that influence vaccine acceptance.

Entering the 21st century, people gradually adapt to the idea that there is no better place than the Internet to find and share information they need at the fingertips, including health information. According to a study conducted by the KRC Research (https://www.webershandwick.com/wp-content/uploads/2018/11/Healthcare-Info-Search-Report.pdf), 81% Americans seek for healthcare information online in 2018, with some indicating that they rely more on the Internet than physicians in search of health advice. In the context of vaccine information acquisition and

sharing, the overdependence on online information is alarming. When examining online antivaccine websites, Kata [16] found that a combination of tactics like shifting hypotheses and censorship makes antivaccine claims highly convincing, thus effectively spreading fear and suspicion towards vaccines among Internet users. As social media platforms become prevalent, more and more researchers utilize user-generated contents to understand the public opinions on vaccines as well as their associations with vaccine coverage. For example, with the use of Twitter data, Dunn et al. [17] found a high correlation between the social media information exposure and the state-level HPV vaccine coverage in the U.S., which can explain the differences in coverage that are unexplainable by socioeconomic factors like education and insurance. Further, in a recent study also focusing on the HPV vaccine, Lama et al. [18] revealed that political debates are the most frequent topics in all discussions related to HPV vaccines on Reddit, in comparison with conversations on general mainstream media which usually link HPV vaccines to sexual activities. In terms of validity, a study which applied natural language processing methods to Twitter data proved that in the case of understanding vaccine refusal, "the strengths of social media data may greatly outweigh their weakness" [19], which aligns with the consensus of top methodologists that the most reliable research comprises mixed methods and data sources [20].

Since the outbreak of the COVID-19 pandemic, several studies have already conducted surveys to reflect the public health concerns regarding COVID-19 vaccines (e.g., Pogue et al. [21]). Nevertheless, none of the aforementioned studies analyzed Reddit discourse about COVID-19 vaccines and the evolution of underlying concerns implied by the change of topics on this issue. Therefore, we believe that our interdisciplinary approach with the use of Reddit data would be a potent complement to this study area. To the best of our knowledge, there is no existing published study conducted under similar conditions. We believe our findings could fill in the gaps currently existing in the understanding of different communities' concerns regarding COVID-19 vaccines, thus offering some useful insights for the design of communication and immunization strategies.

3. Materials and Methods

Reddit consists of individual communities or subgroups that differ in topics (i.e., r/Coronavirus, r/worldnews), which are called subreddits. Reddit users can create original posts, which are referred to as submissions, on a particular subreddit, and comment under submissions. If any user replies to a comment and carries on the conversation under that comment, all the comments together form a comment forest, where it is common that the topics of a comment forest deviates from the original topic discussed in the submission or the top comment. Compared to submissions, which are mostly news, questions, and lengthy stories, comments under submissions are better sources in the light of discourse characterizing, so we decide to focus on top comments under the submissions about COVID-19 vaccines.

To collect the comments, we employ a Reddit API Wrapper called PRAW (https://praw.readthedocs.io/en/latest/), which is a Python package extensively used in Reddit-related studies (e.g., Buntain and Golbeck [22]). We crawl the comments through the keyword-searching function in the package, so the comments collected must contain at least two keywords: one from the list ["vaccine", "vaccines", "vaccinated", "vaccination", "vacinne", "vacine"] plus one from the list ["covid", "covid-19", "coronavirus", "pandemic", "immunization"]. With the combination of these two keyword lists, we are able to crawl only comments regarding COVID-19 vaccines while excluding those that are related to either other vaccines (i.e., HPV vaccines, flu shots) or vaccine-unrelated COVID-19 topics. Unlike some studies that collected data from one or two particular subreddits (e.g., Gozzi et al. [23]), we crawl comments that meet the conditions in a sitewide basis. For each comment, we download its subreddit, comment id, comment created time, comment author, and comment body. All of the data are visible and accessible to the public. Taking into account the fact that the COVID-19 pandemic did not expand to a worldwide scale until March—the U.S. reported its first coronavirus death on February 29, 2020 (https://www.cdc.gov/media/releases/2020/s0229-COVID-19-first-death.html), and the WHO declared the pandemic on March 11, 2020—we discard comments which predated March 1, 2020. Duplicate comments posted in the same or different subreddits and bots are also pruned. As a result, we had 172,091 comments from 6,466 subreddits that were generated by 107,522 unique users, spanning from March 1, 2020, to December 15, 2020. Table 1 lists the top eight subreddits ranked by the number of comments, along with their general attributes. The total numbers of comments and authors of these 8 subreddits constitute over 1/3 of the entire comments and authors in all the subreddits.

3.1. LDA. Based on the evidence presented in a recent study [24], LDA not only is the most popular topic modeling method to date but also achieves better performance in a short-text context among many other topic modeling methods such as the Latent Semantic Analysis (LSA). Moreover, the study also proves that LDA can produce higher-quality and more coherent topics than other unsupervised topic modeling algorithms. Therefore, LDA is chosen for topic extraction in our study. Before feeding the comments into the LDA model, we remove punctuations from the comments and conduct lemmatization. We also process the data by using the directory of "English" stopwords from the Natural Language Toolkit (NLTK) (https://www.nltk.org/book/ch02.html) with an extended list of words ["vaccine", "vaccines", "vaccinated", "covid", "coronavirus", "virus", "us", "get", "take", "also"], which is aimed at best differentiating possible topics. To determine the most appropriate number of topics, we train LDA models with different numbers of topics using Gensim (https://radimrehurek.com/gensim/) and graph their respective coherence scores. The highest coherence score (0.540) is reached when the number of topics is set to 38. However, after manually inspecting the keywords of the 38 topics, we find that many topics are overlapping

Table 1: Top eight subreddits with the most number of comments.

Subreddits	Num. of comments	Num. of authors	Average comments per author (Std)	Median	Max	Min
r/Coronavirus	17,263	9,849	1.75 (3.18)	1	114	1
r/worldnews	7,985	5,782	1.38 (8.93)	1	676	1
r/conspiracy	7,122	4,159	1.71 (3.02)	1	145	1
r/politics	7,006	5,520	1.27 (1.55)	1	71	1
r/wallstreetbets	6,424	3,982	1.61 (2.14)	1	58	1
r/AskReddit	5,700	4,954	1.15 (1.10)	1	55	1
r/news	3,779	3,123	1.21 (0.79)	1	20	1
r/COVID-19	2,378	1,169	2.03 (4.12)	1	64	1
Grand total	57,657	35,798	1.61 (4.78)	1	747	1

with each other. Therefore, we make several attempts by changing the number of topics until a number is found that both produces coherent topics and has a relatively high coherence score, which occurs at 10. The coherence score of the chosen model is 0.511.

3.2. LIWC. We apply LIWC 2015 to capture the psychological information from the collected comments. It reads a given text and counts the percentage of words that reflect different emotions, thinking styles, and social concerns (https://radimrehurek.com/gensim/). Lyu et al. [10] have suggested that sentiment is positively related to the acceptance level of COVID-19 vaccines; therefore, posemo (i.e., positive emotion) and negemo (i.e., negative emotion) categories are selected to measure the sentiment expressed through the comments. To get more nuanced insights into emotions, anx (i.e., anxiety) and anger are included. Health, risk, and death are also included to obtain a broader understanding of the comments. Similar to the methods of Chen et al. [25], we concatenate all the comments of the authors of the same group and apply LIWC 2015. Since the scales of the LIWC categories are not the same, we normalize the LIWC scores in the plots to better illustrate the qualitative differences.

4. Results

4.1. RQ 1 Results. Table 2 itemizes the 10 topics that predominate the comments. According to the keywords under each topic and their probability of occurring in the corresponding topic, we manually designate topic labels to the topics generated by the model, which is a common practice in studies that also use LDA for topic modeling (e.g., Okon et al. [26]). Note that every comment may belong to more than one topic, while in this case, we categorize each comment to only its dominant topic in order to calculate the proportion of comments for each topic. The first row indicates that among all the comments, 12.44% of them belong to the topic "skeptical/aggressive remarks," and the top 10 keywords associated with this topic are listed in a descending order of contribution. Not only does the first topic occupy the highest percent of comments, but also the percentage difference between it and the second topic "clinical trials/research/testing" is larger (1.22%) than the difference between any other two consecutive topics. The Appendix lists the representative example of comments for each topic, respectively.

As shown in the example of "skeptical/aggressive remarks," one common feature of this type of comments is that they convey conspicuous suspicions or condemnations of certain entities, which could be antivaxxers, vaccine manufacturers, governments, etc. From a longitudinal perspective, Figure 1 manifests that comments of "skeptical/aggressive remarks" predominated the overall discourse related to COVID-19 vaccines from June to December except for November (6 out of 10 months). Even though comments of "governments/big companies" supplanted the dominant position of comments belonging to "skeptical/aggressive remarks" in November, the number of the latter still outstripped the number of comments of any other topics till the end of the collection period. In contrast, comments about "lockdown/spread/cases" originally had the largest proportion among all comments in April and May, but the comments' proportion decreased sharply from May to June and remained as the least discussed topic till the end. In a holistic view, there is a salient spike caused by the increase of the amount of all comments in November, which may be attributed to real-world events carrying prominent importance, such as the 2020 U.S. Presidential Election and BioNTech and Pfizer's announcement that their vaccine is 95% effective.

From a cross-sectional perspective, Figure 2 illustrates the proportions of comments of different topics in the top eight subreddits. Particularly, comments of "skeptical/aggressive remarks" occupy the largest proportion in r/conspiracy; comments of "stock market/sports" have the largest proportion in r/wallstreetbets; comments of "clinical trials/-research/testing" and "symptoms/immune systems" predominate r/COVID-19. It is noteworthy that the dominant topic detected by our LDA model in each subreddit is in accordance with the subreddit's description. For example, in the "About Community" field of r/COVID-19, it reads "In December 2019, SARS-CoV-2, the virus causing the disease COVID-19, emerged in the city of Wuhan, China. This subreddit seeks to facilitate scientific discussion of this global public health threat" (https://www.reddit.com/r/COVID19/). This finding keeps constant in the rest of the subreddits, which proves the robustness of our topic characterizing model. Moreover, the proportion of comments of the same topic varies widely from subreddit to subreddit, which may

Table 2: 10 topics generated by LDA and their associated keywords.

Topics	% of comments	Keywords
Skeptical/aggressive remarks	12.44%	People, mask, wear, make, fuck, anti, thing, literally, shit, stop
Clinical trials/research/testing	11.22%	Test, trial, study, phase, result, dose, effective, testing, early, datum
Life/family/kids	11.18%	Work, live, home, feel, day, life, school, family, child, back
People/vaccine efficacy/risks	10.35%	People, immunity, death, risk, die, rate, population, case, high, number
Governments/big companies	10.16%	Government, make, company, world, country, pandemic, work, money, pay, develop
Symptoms/immune systems	9.38%	Effect, immune, antibody, response, human, system, develop, body, side, cell
Time/long-term effects	9.15%	Year, long, time, month, thing, good, make, term, happen, bad
Stock market/sports	8.96%	Back, market, start, big, year, stock, business, buy, play, hold
Politics/news sources	8.80%	Trump, https, comment, question, article, post, news, source, claim, pandemic
Lockdown/spread/cases	8.36%	Case, lockdown, health, country, spread, hospital, social, state, open, place

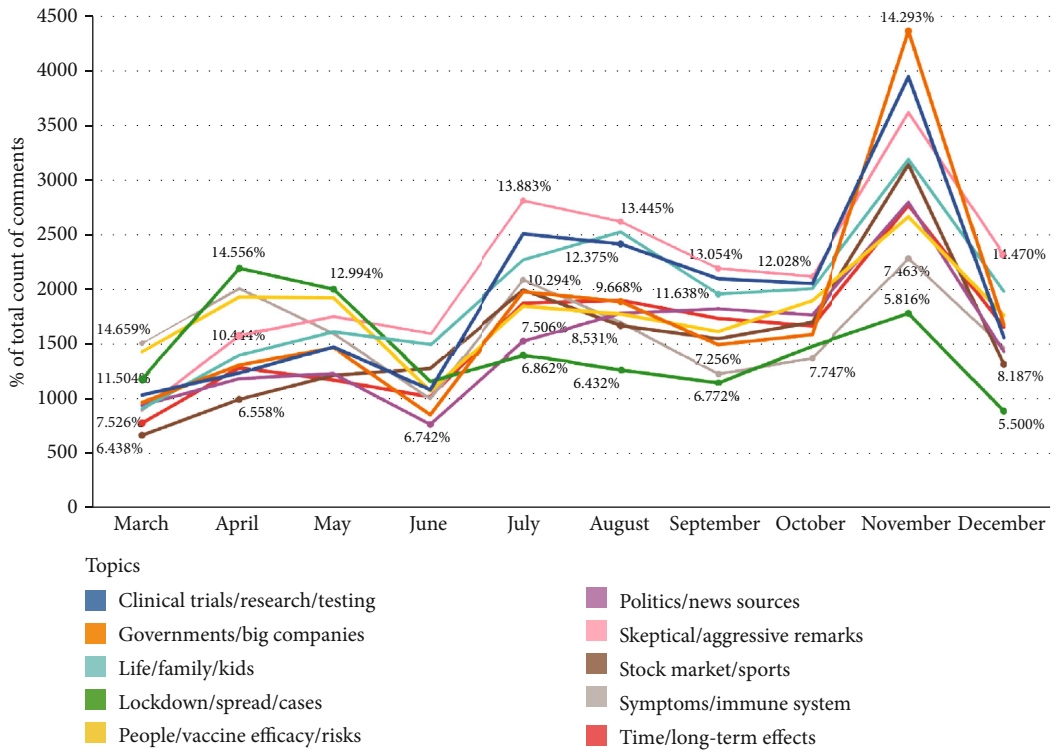

Figure 1: Trends of the proportions of comments of different topics from March 1, 2020 to December 15, 2020.

be caused by two conditions: Reddit users tend to discuss particular topics in particular subreddits, or the fundamental user bases of different subreddits differ and have disparate concerns. In the first case, it is difficult for us to prove whether the comments made by users can represent their major concerns, since the variation may also be caused by the policy of a subreddit; while in the second case, we can prove that the comments made in a specific subreddit can characterize primary concerns of its users, as long as its user base hardly overlaps with that of other subreddits.

4.2. RQ 2 Results. To understand whether the disparity in proportions of comments derive from the differences in subreddits' user bases, we pick three subreddits out of the top eight subreddits: r/Coronavirus, r/worldnews, and r/conspiracy. They are the subreddits that have the most comments among the collected data. In Table 3, we divide them into three pairs and compare each pair's overlapping user bases and their most discussed topics in each subreddit as well as their linguistic profiles.

4.3. r/Coronavirus vs. r/worldnews. This pair shares the most overlapping users (435) among all three pairs. Respectively, the proportions of overlapping users in the user bases of these two subreddits are 4.42% (435 out of 9,849) and 7.52% (435 out of 5,782). In average, each of the 435 users posted four comments in r/Coronavirus and two comments in r/worldnews, which implies that these overlapping users are more active in r/Coronavirus than r/worldnews during the selected time period. In both subreddits, the topic discussed most by

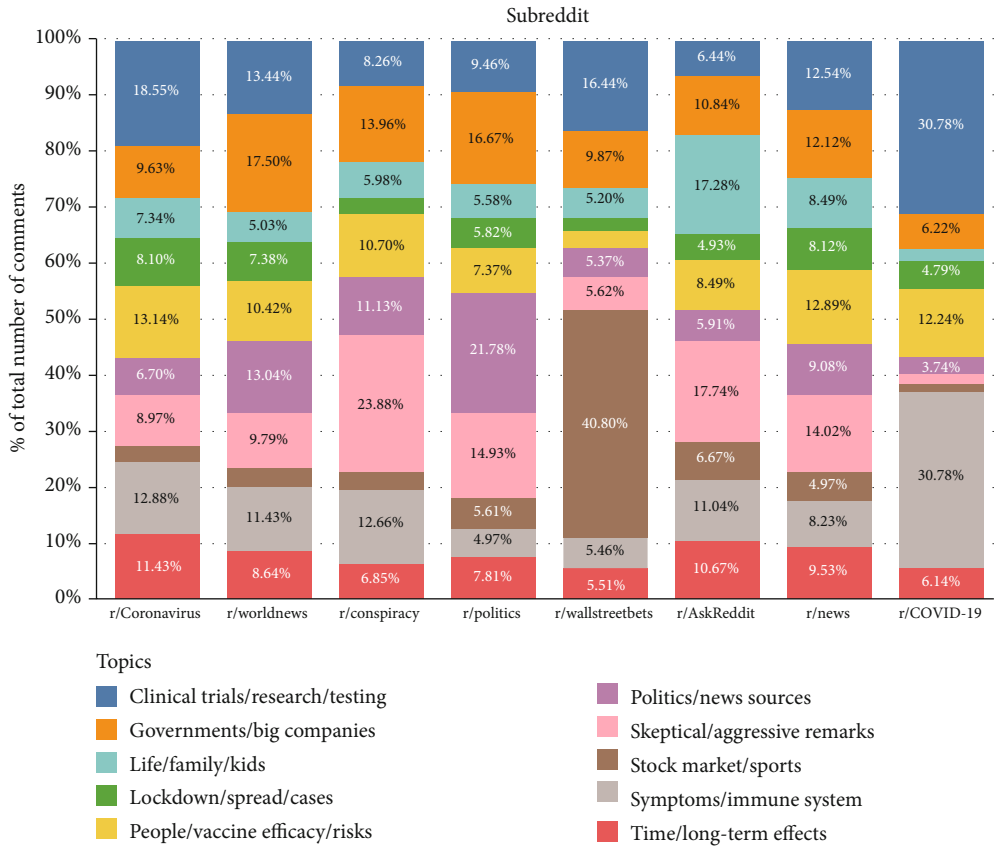

Figure 2: Distributions of comments of different topics across top eight subreddits.

these users is "clinical trials/research/testing," whereas the proportion of comments related to this topic is relatively larger in r/Coronavirus than in r/worldnews, which echoes the findings in the linguistic profiles where the overlapping users pay less attention to the health-related issues in r/worldnews. Conversely, the proportions of comments related to "governments/big companies" and "lockdown/-spread/cases" are both higher in r/worldnews. Interestingly, as shown in Figure 3, the overlapping users express more negative emotions, anxiety, and anger and show more concerns about death-related issues in r/worldnews than r/Coronavirus.

4.4. r/Coronavirus vs. r/conspiracy. This pair shares 104 overlapping users, each of whom posted an average of 2 comments in both subreddits during the selected period of time. Respectively, the overlapping users occupy 1.06% (104 out of 9,849) and 2.50% (104 out of 4,159) of the overall user bases of r/Coronavirus and r/conspiracy. The top five topics that characterize the comments of the overlapping users in these two subreddits are nearly the same, with slight variations in the proportions of comments of each topic. In particular, the overlapping users talked more about "time/long-term effects" in r/Coronavirus, while they made more comments about "skeptical/aggressive remarks" in r/conspiracy. The overlapping users show more anxiety, pay more attention to risks, and talk more about health-related issues in r/Coronavirus, which corresponds to the higher proportion of the comments about "time/long-term effects." In addition, more anger is observed among the overlapping users in r/conspiracy who talked more about "skeptical/aggressive remarks" (Figure 4).

4.5. r/conspiracy vs. r/worldnews. This pair has the least overlapping users (91) among the three pairs. The proportions of overlapping users in the user bases of these two subreddits are 2.19% (91 out of 4,159) and 1.57% (91 out of 5,782). These users made an average of two comments in r/conspiracy and an average of one comment in r/worldnews. Since the difference between the means is statistically significant, it is reasonable to declare that these overlapping users are more active in r/conspiracy than r/worldnews. In terms of the top five occurring topics, although comments related to "skeptical/aggressive remarks" seem to be more frequent in r/conspiracy, the z-test shows no statistically significant difference between the two subreddits' proportions of comments related to any of the listed topics. As for the linguistic profiles, the overlapping users express more positive emotions, focus more on risks, and show more anger in r/worldnews (Figure 5).

In general, r/conspiracy shares a relatively small number of overlapping users with either r/Coronavirus or r/worldnews. Despite some nuances in the proportions of different types of comments, the main topics discussed by overlapping users in disparate subreddits are nearly identical, which refutes the hypothesis that the same users discuss different

Table 3: Characteristics of comment authors who participated in multiple subreddits and most commonly discussed topics by these overlapping authors.

	r/Coronavirus	r/worldnews
Num. of overlapping comment authors	435	435
Num. of total comment authors	1682	717
Mean of Num. of comments	3.87**	1.65**
Top five occurring topics (% of the overall comments)	Clinical trials/research/testing (26.46%)**	Clinical trials/research/testing (20.22%)**
	Symptoms/immune systems (15.10%)	Governments/big companies (17.57%)†
	Governments/big companies (14.80%)†	Symptoms/immune systems (16.04%)
	People/vaccine efficacy/risks (12.25%)	People/vaccine efficacy/risks (12.41%)
	Time/long-term effects (9.39%)	Lockdown/spread/cases (8.09%)
	r/Coronavirus	r/conspiracy
Num. of overlapping comment authors	104	104
Num. of total comment authors	183	220
Mean of Num. of comments	1.76	2.12
Top five occurring topics (% of the overall comments)	People/vaccine efficacy/risks (17.49%)	People/vaccine efficacy/risks (15.45%)
	Governments/big companies (14.75%)	Governments/big companies (15.45%)
	Symptoms/immune systems (13.66%)	Symptoms/immune systems (14.55%)
	Clinical trials/research/testing (13.11%)	Skeptical/aggressive remarks (14.09%)†
	Time/long-term effects (11.48%)†	Politics/news sources (10.91%)
	r/conspiracy	r/worldnews
Num. of overlapping comment authors	91	91
Num. of total comment authors	192	129
Mean of Num. of comments	2.11**	1.42**
Top five occurring topics (% of the overall comments)	Skeptical/aggressive remarks (16.67%)	Clinical trials/research/testing (14.73%)
	People/vaccine efficacy/risks (15.10%)	People/vaccine efficacy/risks (14.73%)
	Symptoms/immune systems (14.58%)	Symptoms/immune systems (13.95%)
	Clinical trials/research/testing (13.02%)	Skeptical/aggressive remarks (13.95%)
	Governments/big companies (12.50%)	Governments/big companies (13.17%)

†p value < 0.1; *p value < 0.05; **p value < 0.01.

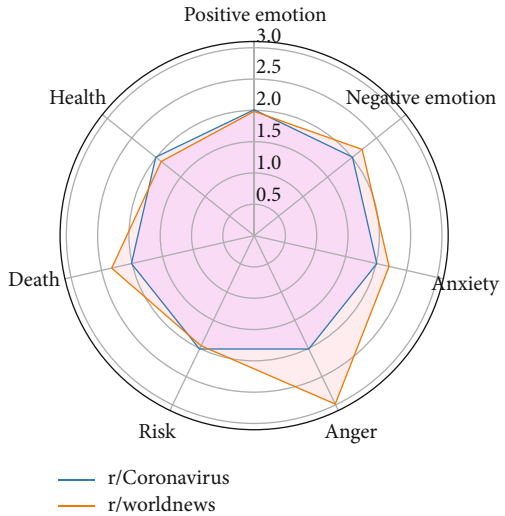

Figure 3: Linguistic profiles for the comments of the overlapping users of r/Coronavirus and r/worldnews (note: the scores have been normalized).

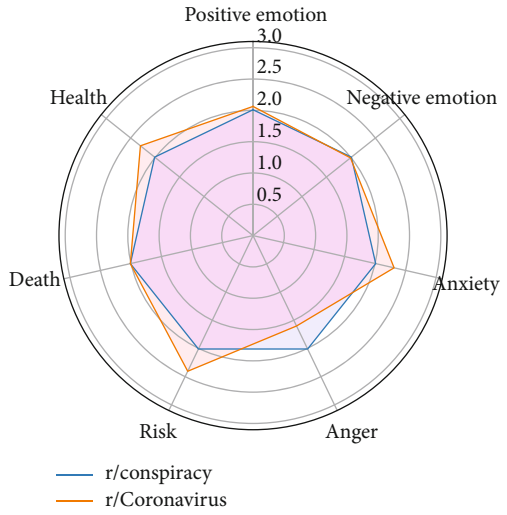

Figure 4: Linguistic profiles for the comments of the overlapping users of r/Coronavirus and r/conspiracy (note: the scores have been normalized).

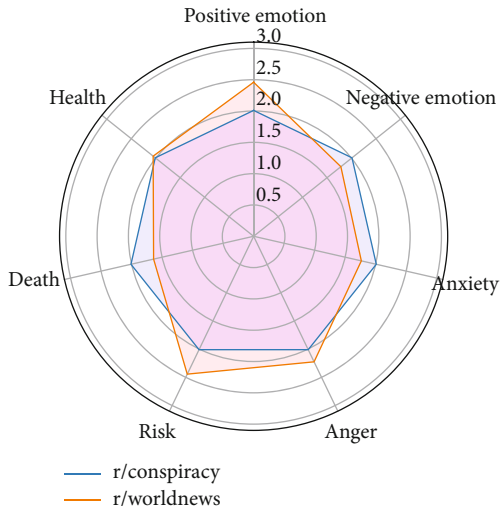

Figure 5: Linguistic profiles for the comments of the overlapping users of r/conspiracy and r/worldnews (note: the scores have been normalized).

topics based on the features/policies of specific subreddits and thus contribute to the variations in different comments' proportions to a large extent. In other words, the impact of the overlapping users is modest. Therefore, it is valid to infer that the variations in topic proportions across subreddits are mainly attributed to different user bases, which is confirmed by the small percentage of overlapping users in the overall user base of each subreddit. A further qualitative analysis of the linguistic profiles of the comments of overlapping users is conducted. The differences in the linguistic profiles, combined with topics, have revealed more insights into the user bases. It is especially worth noting that r/conspiracy exceeds r/Coronavirus in anger and r/worldnews in death, anxiety, and negative emotion, respectively. Instead of what the word "conspiracy" suggests about the subreddit, we find that comments from this subreddit vary in stances on COVID-19 vaccines: some of them question the intents of governments/big companies with suspicions and malignity, while others condemn those who are suspected as antivaxxers. One common linguistic feature of these comments is the dense use of offensive/profane words, which also contributes to the dominant proportion of "skeptical/aggressive remarks" in the subreddit. To reveal the complexity of opinions, we include representative examples from r/conspiracy in the Appendix.

5. Discussion

By examining what specific topics characterize the discussions of COVID-19 vaccine on Reddit, we find that the proportion of the comments about "skeptical/aggressive remarks" outweighed that of other topic during the pandemic. Moreover, these topics vary across time and subreddits, suggesting differences in subreddits' user bases. We further pick three subreddits out of the top eight: r/Coronavirus, r/worldnews, and r/conspiracy, and find that the variations in topic proportions across subreddits are mainly related to the thematic and linguistic differences among the overlapping users.

5.1. Implications. As the first work to characterize discourse related to COVID-19 vaccines on Reddit, our study has three significant implications. First and foremost, aside from Twitter [27], Reddit has served as a "hotbed" for conspiracy theories and disinformation since the outbreak of the pandemic. The facts that r/conspiracy has gained the third most comments and that the sitewide conversations have been dominated by those belonging to "skeptical/aggressive remarks" for six months are simply shocking. Nevertheless, this finding is in accordance with the claim that the 21st century is "the golden age of anti-vaccine conspiracies", made by Stein [28]. Second, despite the small number of overlapping users between subreddits, each of the most active subreddits has its own user bases. Hence, reliable news posted in r/Coronavirus would not draw attention from users in r/conspiracy, while the latter may be the group of people who need reliable information sources the most. Lastly, even on a single social media platform like Reddit, the concerns related to COVID-19 vaccines vary from subcommunity to subcommunity, not to mention more popular social media platforms like Facebook and Twitter. These various concerns imply different communication preferences [29]. Therefore, in real practice, we must consider the characteristics of different groups and implement communication strategies targeting at specific groups' needs. *Speak the same language* is the weapon for us to overcome the overabundance of mis- and disinformation in times of crisis.

In terms of tackling the challenges suggested by the findings, active actions and collaborations from multiple parties are needed. As stated by policy advisors Wardle and Singerman, "we need responses that acknowledge the complexity of defining misinformation, of relying on scientific consensus, and of acknowledging the power of narratives" [30]. Our study contributes to the first stage by quantifying multifaceted discussions regarding COVID-19 vaccines and further disentangle them using NLP techniques. Our study also sheds light on the disagreements among different subgroups by providing concrete examples and comparing them on a user—as well as thematic—level. Based on what we find and other pertinent studies, social media platforms and policy makers should obtain a more comprehensive landscape of online discussions and thus be able to create more effective strategies that build the public's consensus and trust in vaccinations.

5.2. Limitations. Admittedly, our study is not free of limitations. Although our model achieves a high coherence score in characterizing the collected comments, the generated topics can only represent the concerns of users who posted top comments that meet our selection standards (keyword searching, etc.). There may be some opinions or concerns underlying comments that were not taken into account in our current analysis, so in order to obtain a more precise picture of thematic characteristics of the discourse related to COVID-19 vaccines on Reddit, we should amplify the scale of data in future works. In addition, since it is impossible to

know Reddit users' demographics, our findings simply imply the discourse characteristics of the general Reddit users instead of the users from a certain region or socioeconomic level. From the perspective of designing communication or immunization programs, more information is needed to target the needs of specific groups. Moreover, we did not incorporate the interactions between Reddit users (i.e., upvotes and downvotes) in this work, although these are elements that may help us better understand the dynamics of pertaining conversations—this can be done in future works.

6. Conclusion

In this study, we characterize the Reddit discourse related to COVID-19 vaccines through a combination of computational and qualitative methods. Specifically, we employ an LDA model to categorize top Reddit comments from March 1, 2020, to December 15, 2020, under 10 topics; analyze how the number of each type of comments changed over time; and examine these comments' proportions across eight different subreddits—we detect discrepancies from both longitudinal and cross-sectional perspectives. Thereafter, we conduct a more careful investigation into the thematic and linguistic variations between subreddits by comparing three most active subreddits. With the use of statistical tests, we confirm that although there are overlapping users between these subreddits, the scale of them is quite smaller than that of each subreddit's own user base, thus proving the hypothesis that thematic variations among subreddits' conversations mainly result from differences in user bases. This finding also suggests the need of taking care of different subgroups' communication preferences within a large social media platform. Furthermore, we employ LIWC and find that there are also differences among the overlapping users across emotions, personal concerns, and attention. As information on social media platforms puts greater influence on people's vaccination behaviors [16], we believe the findings of our study can provide policy makers with useful insights in the process of designing effective communication and immunization programs.

Appendix

A. Example Comments under the 10 Topics

Topic 1: clinical trials/research/testing. "Normally it takes a long time to enroll enough people. It also takes a while for people to become infected with the thing that you're studying. We don't have that problem in a pandemic where almost 5 million people every week is infected with COVID-19. It's the same process, but it doesn't take as long to get results. […]".

Topic 2: governments/big companies. "As someone who is vaccinated and whose children are vaccinated I won't touch this vaccine with a 10 foot pole. When its so safe that the manufacturers demand blanket protection from any Liability from side effects and the countries governments give it to them that should already be a huge red flag. Pfizer: our vaccine is super safe and will help stop Covid-19. Also Pfizer: we're going to need complete immunity from liability if our vaccine turns out to be a complete fiasco. […]".

Topic 3: life/family/kids. "It really concerns me for my children more than myself. I don't want to give them a chemical during their childhood that may fuck them up medically later in life. We homeschool them. We don't go out unnecessarily, we wear masks, we social distance, etc. I'm not against ever taking a covid vaccine but I will definitely be waiting before taking it or giving it to my kids. […]".

Topic 4: lockdown/spread/cases. "[…] My state | <state name> | has been completely free of any community transmission for over seven months and life has basically returned to normal. The only cases have been people returning from overseas who have to undergo a mandatory quarantine for two weeks when they arrive".

Topic 5: people/vaccine efficacy/risks. "Something like 12% of the population have Hashimoto's. If the condition was a significant risk factor with a vaccine, it WOULD have shown up in the clinical trials. Even under the worst possible assumptions, the risks from Covid would be literally a thousand times higher than from a vaccine. Anyone who needs any convincing for taking the vaccine should visit the Covid long-hauler board".

Topic 6: politics/news sources. "Cool. In the meantime you could actually read about vaccine distribution, here's some useful links: <source links>".

Topic 7: skeptical/aggressive remarks. "[…] See you out in the real world very very soon vaccine is coming, wooooohoooo! Can't wait to burn my mask in dumpster fire! Vaccine vaccine vaccine!! Goodbye covid!!!!! Even <name> and <name> are like who gives a fukkkkk!"

Topic 8: stock market/sports. "The fact so many stocks have risen to higher valuations then the Pre-March covid crash is just laughable. All these sad people will be left holding bags once the Rich take a massive dump on all of them. Everything is essentially at an ATH, Over valued and propt up by Vaccine optimism/Governments Stimulus. The rug will be pulled out from under the average persons feet trying to invest for retirement".

Topic 9: symptoms/immune systems. "I am O- and have had Covid. I had moderate symptoms for 36-48 hours (low fever, night sweats, chills, body aches, diarrhea). I did not have respiratory issues and did not develop a cough. I'm an extra complicated case, however, as within the past six years, I have underwent 2.5 years of chemo for acute lymphoblastic leukemia (ALL). I am quite terrified of how Covid or the vaccine might affect the ALL, which is currently in remission. Thought I'd throw this out there in case anyone smarter than me has any insights".

Topic 10: time/long-term effects. "[…] Like you said, it has been less than a year, we don't known jack shit yet, but they want to inject us all with rushed vaccines we don't even know the effectiveness for or how long that effectiveness will last. And you're okay with that?"

B. Example Comments from r/Conspiracy

"Vaccines wouldn't have been necessary, and this fucking pandemic wouldn't still be such a major issue if this

ridiculous administration hadn't dissolved the group specifically designed to combat this process and if the population at large had followed simple fucking suggestions. But they can't and here we are… saying 'No' to a fucking virus."

"Somebody who's anti vaxx wouldn't accept a COVID vaccine no matter what the top health officials said. They lied to us about aids and now they are lying about COVID. Do not take the test do not take the vaccine it's the aids epidemic all over again!!!! vaccines are cool. I'm probably been vaccinated more times than anyone on this sub or even the provax subs, and I even got a vaccine for anthrax for no reason once and it hurt like hell. It's like getting hit in the arm by a boxer and doesn't go away for a couple days."

"Anyway, like I said, vaccines are fine, but in this case they're not. It's clear that there could be a way to treat COVID 19 without a worldwide mandatory vacccine, yet any research regarding that is being heavily suppressed and unfairly criticized by the mainstream media and Dr. Fauci. Does that seem like good science? Not really."

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Authors' Contributions

W. Wu, H. Lyu, and J. Luo conceived the idea and designed the experiments. W. Wu and H. Lyu conducted the experiments. W. Wu wrote the majority of manuscript. H. Lyu wrote part of the manuscript. J. Luo supervised the study and reviewed the manuscript.

Acknowledgments

This research was supported in part by a University of Rochester Research Award and NIH grant RF1AG063811-01S2.

References

[1] N. Fernandes, "Economic effects of coronavirus outbreak (covid-19) on the world economy," *SSRN Electronic Journal*, 2020.

[2] A. Atalan, "Is the lockdown important to prevent the COVID-19 pandemic? Effects on psychology, environment and economy-perspective," *Annals of Medicine and Surgery*, vol. 56, pp. 38–42, 2020.

[3] M. Lipsitch and N. E. Dean, "Understanding COVID-19 vaccine efficacy," *Science*, vol. 370, no. 6518, pp. 763–765, 2020.

[4] J. V. Lazarus, S. C. Ratzan, A. Palayew et al., "A global survey of potential acceptance of a COVID-19 vaccine," *Nature Medicine*, vol. 27, pp. 225–228, 2020.

[5] P. L. Reiter, M. L. Pennell, and M. L. Katz, "Acceptability of a COVID-19 vaccine among adults in the united states: How many people would get vaccinated?," *Vaccine*, vol. 38, no. 42, pp. 6500–6507, 2020.

[6] R. D. Goldman, T. D. Yan, M. Seiler et al., "Caregiver willingness to vaccinate their children against COVID-19: cross sectional survey," *Vaccine*, vol. 38, no. 48, pp. 7668–7673, 2020.

[7] P. S. Brenner and J. DeLamater, "Lies, damned lies, and survey self-reports? Identity as a cause of measurement bias," *Social Psychology Quarterly*, vol. 79, no. 4, pp. 333–354, 2016.

[8] Z. Jin, J. Cao, H. Guo, Y. Zhang, Y. Wang, and J. Luo, "Detection and analysis of 2016 US presidential election related rumors on Twitter," in *Social, Cultural, and Behavioral Modeling*, pp. 14–24, Springer, 2017.

[9] X. Lu, L. Chen, J. Yuan et al., "User perceptions of different electronic cigarette Flavors on social media: observational study," *Journal of Medical Internet Research*, vol. 22, no. 6, article e17280, 2020.

[10] H. Lyu, J. Wang, W. Wu et al., "Social media study of public opinions on potential COVID-19 vaccines: informing dissent, disparities, and dissemination," 2020, https://arxiv.org/abs/2012.02165.

[11] D. K. Kilgo, Y. M. M. Ng, M. J. Riedl, and I. Lacasa-Mas, "Reddit's veil of anonymity: Predictors of engagement and participation in media environments with hostile reputations," *Social Media + Society*, vol. 4, no. 4, 2018.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[13] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: LIWC 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, article 2001, 2001.

[14] H. J. Larson, C. Jarrett, E. Eckersberger, D. M. Smith, and P. Paterson, "Understanding vaccine hesitancy around vaccines and vaccination from a global perspective: a systematic review of published literature, 2007-2012," *Vaccine*, vol. 32, no. 19, pp. 2150–2159, 2014.

[15] A. de Figueiredo, I. G. Johnston, D. M. D. Smith, S. Agarwal, H. J. Larson, and N. S. Jones, "Forecasted trends in vaccination coverage and correlations with socioeconomic factors: a global time-series analysis over 30 years," *The Lancet Global Health*, vol. 4, no. 10, pp. e726–e735, 2016.

[16] A. Kata, "Anti-vaccine activists, Web 2.0, and the postmodern paradigm - an overview of tactics and tropes used online by the anti-vaccination movement," *Vaccine*, vol. 30, no. 25, pp. 3778–3789, 2012.

[17] A. G. Dunn, D. Surian, J. Leask, A. Dey, K. D. Mandl, and E. Coiera, "Mapping information exposure on social media to explain differences in HPV vaccine coverage in the United States," *Vaccine*, vol. 35, no. 23, pp. 3033–3040, 2017.

[18] Y. Lama, D. Hu, A. Jamison, S. C. Quinn, and D. A. Broniatowski, "Characterizing trends in human papillomavirus vaccine discourse on reddit (2007-2015): an observational study," *JMIR Public Health and Surveillance*, vol. 5, no. 1, article e12480, 2019.

[19] M. Dredze, D. A. Broniatowski, M. C. Smith, and K. M. Hilyard, "Understanding vaccine refusal: why we need social media now," *American Journal of Preventive Medicine*, vol. 50, no. 4, pp. 550–552, 2016.

[20] A. Adams, S. Soumerai, J. Lomas, and D. Ross-Degnan, "Evidence of self-report bias in assessing adherence to guidelines," *International Journal for Quality in Health Care*, vol. 11, no. 3, pp. 187–192, 1999.

[21] K. Pogue, J. L. Jensen, C. K. Stancil et al., "Influences on attitudes regarding potential COVID-19 vaccination in the united states," *Vaccines*, vol. 8, no. 4, p. 582, 2020.

[22] C. Buntain and J. Golbeck, "Identifying social roles in reddit using network structure," in *Proceedings of the 23rd international conference on world wide web*, pp. 615–620, New York, NY, USA, 2014.

[23] N. Gozzi, M. Tizzani, M. Starnini et al., "Collective response to media coverage of the COVID-19 pandemic on reddit and wikipedia: mixed-methods analysis," *Journal of Medical Internet Research*, vol. 22, no. 10, article e21597, 2020.

[24] R. Albalawi, T. H. Yeap, and M. Benyoucef, "Using topic modeling methods for short-text data: a comparative analysis," *Frontiers in Artificial Intelligence*, vol. 3, p. 42, 2020.

[25] L. Chen, H. Lyu, T. Yang, Y. Wang, and J. Luo, "Fine-grained analysis of the use of neutral and controversial terms for COVID-19 on social media," in *Social, Cultural, and Behavioral Modeling. SBP-BRiMS 2021*, Lecture Notes in Computer Science, R. Thomson, M. N. Hussain, C. Dancy, and A. Pyke, Eds., pp. 57–67, Springer, Cham, 2021.

[26] E. Okon, V. Rachakonda, H. J. Hong, C. Callison-Burch, and J. B. Lipo, "Natural language processing of Reddit data to evaluate dermatology patient experiences and therapeutics," *Journal of the American Academy of Dermatology*, vol. 83, no. 3, pp. 803–808, 2020.

[27] A. M. Jamison, D. A. Broniatowski, M. Dredze, A. Sangraula, M. C. Smith, and S. C. Quinn, "Not just conspiracy theories: vaccine opponents and pro-ponents add to the COVID-19 `infodemic' on Twitter," *Harvard Kennedy School Misinformation Review*, vol. 1, no. 3, 2020.

[28] R. A. Stein, "The golden age of anti-vaccine conspiracies," *Germs*, vol. 7, no. 4, pp. 168–170, 2017.

[29] M. B. Gilkey, W. A. Calo, M. W. Marciniak, and N. T. Brewer, "Parents who refuse or delay HPV vaccine: differences in vaccination behavior, beliefs, and clinical communication preferences," *Human Vaccines & Immunotherapeutics*, vol. 13, no. 3, pp. 680–686, 2017.

[30] C. Wardle and E. Singerman, "Too little, too late: social media companies' failure to tackle vaccine misinformation poses a real threat," *BMJ*, vol. 372, 2021.