Check for updates

Original Research Article

Health Informatics Journal

Health Informatics Journal
1–18
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14604582221142443
journals.sagepub.com/home/jhi
SAGE

Understanding information needs during COVID-19: A comparison study between an online health community and a Q&A platform

Rachel X Peng and Ryan Yang Wang
Donald P. Bellisario College of Communications, Penn State University, University Park, PA, USA

Abstract

This paper aims at identifying user's information needs on Coronavirus and the differences of user's information needs between the online health community MedHelp and the question-and-answer forum Quora during the COVID-19 global pandemic. We obtained the posts in the sub-community *Coronavirus* on MedHelp (195 posts with 1627 answers) and under the topic of *COVID-19(2019-2020)* on Quora (263 posts with 8401 answers) via web scraping built on Selenium WebDriver. After preprocessing, we conducted topic modeling on both corpora and identified the best topic model for each corpus based on the diagnostic metrics. Leveraging the improved sqrt-cosine similarity measurement, we further compared the topic similarity between these two corpora. This study finds that there are common information needs on both platforms about vaccination and the essential elements of the disease including the onset symptoms, transmission routes, preventive measures, treatment and control of COVID-19. Some unique discussions on MedHelp are about psychological health, and therapeutic management of patients. Users on Quora have special interests of information about the association between vaccine and Luciferase, and attacks on Fauci after email trove released. The work is beneficial for researchers who aim to provide accurate information assistance and build effective online emergence response programs during the pandemic.

Keywords

COVID-19, Information needs, MedHelp, Quora, Text mining, Topic modeling

Corresponding author:
Ryan Yang Wang, Donald P. Bellisario College of Communications, Penn State University, 8 Carnegie Building, University Park, PA 16802, USA.
Email: yuw320@psu.edu

Introduction

The growth of the Internet has led to health information being more accessible in the past decade. Studies have reported more than 70% of Internet users have searched online for health information or have used the Internet for health-related purposes.[1,2] Most frequently people went online to search for condition-specific information, or a particular medical treatment or procedure as well as disease prevention.[1]

The COVID-19 pandemic is a public health and medical emergency on an unprecedented scale that began in 2020. COVID-19 as a pandemic can cause feelings of worry, fear, distress and anxiety.[3] Meanwhile, the knowledge that can be gained from the Internet can have a remarkable impact during pandemics. Previous research has established that appropriate and timely access to quality healthcare information during infectious disease outbreaks could help mitigate public anxiety, develop adequate risk perceptions and make proper health decisions to adopt protective measures.[4,5] Therefore, it is crucial for stakeholders to understand the public's preferences for healthcare information. Knowing what topics the public need information about, could assist with the presentation of healthcare information in a manner and format that suits public needs.

Medical journals have emphasized the growing importance of social media platforms as a valuable tool for the dissemination of disease mitigation strategies.[6] The public has become more reliant on social networking sites to stay informed during a crisis.[7] Instead of focusing on social media platforms which are built for socializing via many-to-many conversation such as Twitter or Facebook,[8–10] this study aims to investigate public health information needs for COVID-19 on online health communities (OHCs) and question-and-answer (Q&A) forums, which are built on one-to-many knowledge-based communication. The main aim of this study is to identify a variety of information needs of public in the course of the pandemic. This work will generate fresh insights to help public health agencies know how to communicate and what to focus on as the world continues to navigate through the COVID-19 outbreak.

Drawing on previous research, this study investigates the following three research questions:

Research Question 1(RQ1): What kinds of health-related information do users discuss in the OHC about COVID-19 pandemic?

Research Question 2(RQ2): What kinds of health-related information do users communicate in the Q&A forum about COVID-19 pandemic?

Research Question 3(RQ3): Do users' information needs on COVID-19 in the OHC differ from the Q&A forum?

Related work

Health information needs during COVID-19 pandemic

Information need is a rather nebulous term, difficult to define, isolate and measure.[11] According to Nicholas,[12] "when definitions of the concept information needs are provided, they are typically vague or highly complicated, and individuals often talk about information needs when they are actually referring to information wants and demands" (p. 9). Within the context of health, Ormandy[13] suggested that a need for information would arise when recognizing the knowledge is insufficient to satisfy a healthcare-related goal. Throughout this paper, the term *information need* will refer to the public's desire for more information during the COVID-19 pandemic that is expressed verbally or in active communication to become better informed about self-care and prevention. We use the concept of *information need* to understand what information people require about ongoing health emergencies.

Several studies have begun to examine user information needs during the coronavirus pandemic. Using entity identification and text analysis, Zhao et al.[14] identified 1496 patients with COVID-19 infections from Wuhan, China, and investigated their health information searching behavior on the Chinese social media platform. They reported that the three most searched topics were access to medical care, isolation, and quarantine guidelines, and offline to online support. Springer et al.[15] used Google Trends data to track the search trends and the patterns of worldwide interest, concerns, and information needs during COVID-19. Wei et al. identified 15 categories of questions about COVID-19 across 13 data sources, and the most asked questions are about transmission, prevention, and societal effects of COVID.[16] So far, however, there are no analyses mining the health information needs for COVID-19 on OHC and Q&A websites in the United States. This study sets out to extend previous work by investigating and comparing information needs on an OHC website MedHelp and a Q&A website Quora during the COVID-19 pandemic.

Health information needs on online health communities

OHCs are patient-led sites that focus exclusively on health-related topics. OHCs exist for a wide range of diseases and health issues, from cancer support groups to simple calorie counter forums. Compared to other health-related sites which only allow users to retrieve information, OHCs allow for communication between multiple people.

Scholars have made significant progress on the discoveries of discussion topics in OHCs. Chen[17] used the k-means algorithm to cluster the discussion content from three cancer-related OHCs and unearthed a set of common topics in each: support, treatment experience, disease and medication management. Park and Park[18] investigated the cancer-related information needs among Korean Americans by collecting posts from MissyUSA, one of the largest online communities among Korean in the USA. They identified the most discussed medical topics which are treatment, diagnosis and symptom. To understand the potential information needs among patients with physiological and psychological diseases, Liu et al.[19] used topic modeling and sentiment analysis to study the differences in topics and emotions expressed by the two groups of patients in OHCs. It was shown that people with physical illness have a high degree of attention to the medical treatment, while people with mental illness are actively involved in seeking emotional support in the community.

Health information needs on question-and-answer forums

Question-and-answer (Q&A) websites are knowledge-sharing platforms for asking explicit questions as well as posting answers. Health has been identified as a major domain to observe user interactions and identify user needs in Q&A website research.[20] These websites provide a venue for consumers to seek experiential information from other users for a quick solution to their healthcare concerns.[21] It has previously been observed that Q&A platforms offer a valuable opportunity to better understand users' information needs and concerns about various health issues.[22]

Text mining has been introduced as a useful method for studying the health information needs in Q&A forums from different perspectives. Oh et al.[23] collected cancer related questions posted on Yahoo Answers and investigated cancer-related topic categories by text mining techniques to reveal users' multidimensional information needs. In a study investigating consumer information needs on dietary supplements, Rizvi et al.[24] retrieved a total of 2,820,179 questions and corresponding answers from Yahoo Answer. By implementing an unsupervised topic modeling method, they found that the most sought information by users are "use and adverse effects", "product-related", and "healthy lifestyle". Zhao et al.[14] analyzed 10,862 depression-related posts on the platform named Zhihu, which is the largest Chinese social Q&A platform. By combining LDA and manual methods, the results showed that the users who sought help for depression pay more attention on the information linked to depression symptoms and social activities.

Method

Data collection

In this study, two typical platforms, the OHC website MedHelp and the Q&A website Quora, were chosen as the comparative cases for data collection. MedHelp is one of the earliest and most popular online forums which attracts more than 12 million users browsing monthly.[25] Quora, one of the largest and most popular Q&A forums in America, has attracted over 300 million monthly active users.[26] A key area where Quora shines is health care. All the data collected are in English and mainly for American users based on their website traffic.[1]

Two web scraping scripts written in R 4.0.5 based on Selenium WebDriver[27] were utilized to obtain all posts (each post consisted of one question and its following comments) in the subcommunity *Coronavirus* on MedHelp;[28] and in the topic of *COVID-19(2019-2020)* on Quora.[29] The data collection was finished on 16 October 2021. Given that short answers might not contain meaningful information for our further analysis, we decided to remove all the answers of less than 5 words. Meanwhile, we have also removed all the URLs, punctuations, numbers, and English stopwords, and performed lowercasing and stemming during the data preprocessing following the common practices. All the data collected in the study is readily available to the public and there is no direct interaction with participants during the data collection process. We have also complied with the privacy policy on each online forum respectively, as no personal information on the user-level was obtained or stored at our end and there is no possible way to link a record with a particular individual.[30] Additionally, we have consulted with an Institutional Review Board analyst at our institute to confirm that no ethical approval is required in the current study. For the process and the codes of web scraping, please refer to the Supplemental Material.

Considering that question may not be specific enough with the wording whereas answers contain more meaningful information which is important to the topic modeling, each post we scraped is consisted of the question and its associated comments. We aggregated each question and the corresponding answers/comments into one *post* for further analysis. Table 1 shows the descriptive

Table 1. The descriptive statistics of MedHelp and Quora corpus.

Online forum	MedHelp[a]	Quora[b]
Number of posts	195	263
Number of answers	1627	8401
Word counts of the whole post	208,742	625,155
Number of terms	3515	6784
Number of tokens	46,994	111,799

[a] 4149 of 7664 terms (4149 of 51,143 tokens) were removed due to frequency.
[b] 7694 of 14,478 terms (7694 of 119,493 tokens) were removed due to frequency.

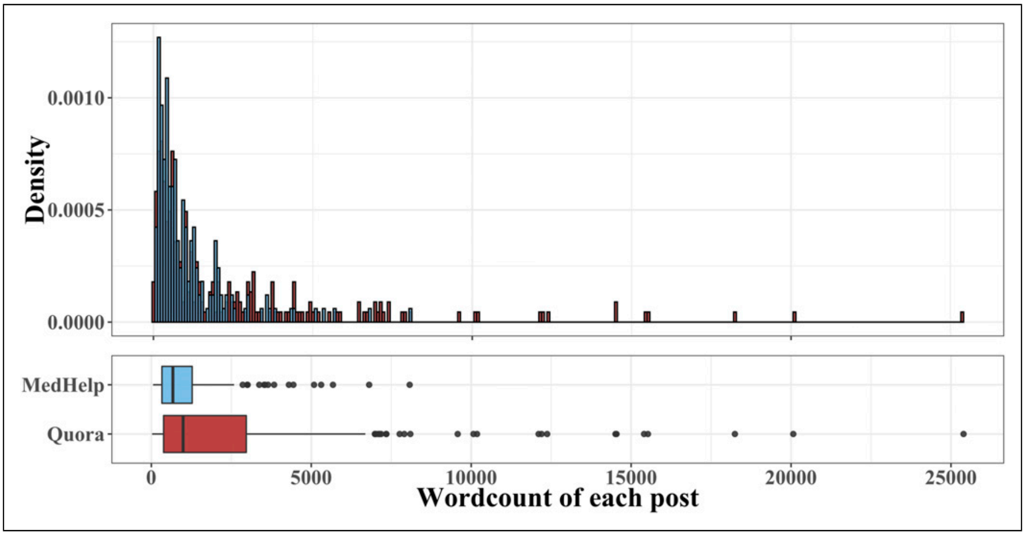

Figure 1. The wordcount distribution of MedHelp and Quora corpus.

statistics of two corpora via the data collection process. There were less questions and less answers in the MedHelp corpus compared to the Quora corpus, and the average post length of MedHelp (1070.47) was also shorter than its Quora counterpart (2377.02). The distribution of the word count for each corpus is shown in Figure 1.

Data analysis

Taking the posts from these two corpora as the unit of analysis, we conducted an in-depth text analysis to understand public health information needs during the COVID-19 pandemic. First, after collecting and cleaning the textual data, we used topic modeling to analyze these two corpora separately to uncover the underlying themes. Second, the differences in users' health information needs between the two sites were investigated through the topic similarities (i.e. improve sqrt-cosine similarity). Our data analysis road map is presented as Figure 2.

Topic modeling. We are using structural topic modeling (STM) without structural metadata to perform the topic modeling via the *stm* R package.[31] A STM without a covariate renders a similar result of Correlated topic modeling (CTM).[2] Correlated topic modeling (CTM) is an unsupervised machine learning algorithm of text mining deriving from Latent Dirichlet allocation model (LDA). LDA renders *K* clusters of co-occurring terms (or the topics) with a bag-of-words approach, which assumes the order of terms and the document in the corpus is neglected in the process. Meanwhile, topics are uncorrelated with each other. CTM is performed as an extension of LDA while allowing the correlations among the latent topics.[31] Moreover, CTM also adapts a logistic normal distribution rather than Dirichlet distribution like LDA does, in order to better adjust the covariance structure among all the topics, which outperforms the regular LDA.[31] The main reason we chose the *stm* package here, is its utilization of elements of SAGE and DMR topic model. Meanwhile, its flexible *searchK* function and the model diagnostic metrics (i.e. semantic coherence, exclusivity, residuals,

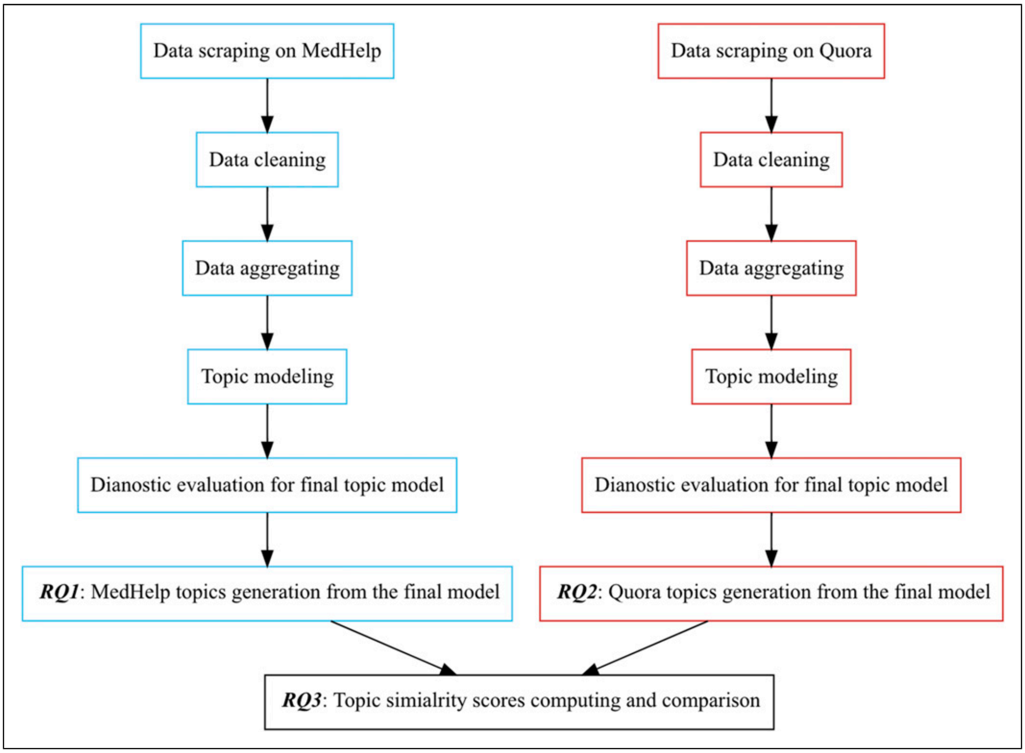
Data scraping on MedHelp
Data cleaning
Data aggregating
Topic modeling
Dianostic evaluation for final topic model
RQ1: MedHelp topics generation from the final model
Data scraping on Quora
Data cleaning
Data aggregating
Topic modeling
Dianostic evaluation for final topic model
RQ2: Quora topics generation from the final model
RQ3: Topic simialrity scores computing and comparison

Figure 2. Data analysis road map.

lower bound, and held-out likelihood), helps us greatly to identify the optimal K of the topic modeling for the MedHelp and Quora corpus.

Topic similarity. Cosine similarity is one of the common methods for comparing the text similarity between documents or two non-zero vectors based on Euclidean distance via a bag-of-words model.[32] Specifically, cosine similarity based on Euclidean distance could be defined as

$$cos(A,\, B) = \frac{A^{*}B}{\|A\|\,\|B\|} = \sum_{i=1}^{n} \frac{A_i{}^{*}\, B_i}{\sqrt{\sum_{i=1}^{n} A_i^2{}^{*}}\, \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

where A and B represent the vectors of word frequencies in the two documents to be compared.

However, Euclidean distance might not be a good metric for dealing with probability, such as the topic-word distribution generated from an LDA or CTM topic modeling.[33] Instead, Hellinger distance with probability-based approaches may be a preferable choice, especially when handling high-dimensional data.[34] As an attempt for a better solution, Sohangir and Wang[35] propose an Improved sqrt-cosine similarity measurement (ISC) based on equation (1) and the sqrt-cosine similarity.[34] Sohangir and Wang contend that the ISC approach outperforms other similarity measurement through their comparative experiments:

$$ISC(A, B) = \frac{\sum_{i=1}^{n} \sqrt{A_i * B_i}}{\sqrt{\sum_{i=1}^{n} A_i *} \sqrt{\sum_{i=1}^{n} B_i}} \tag{2}$$

Similarly, the similarity score for each pair of documents with the ISC approach also ranges from 0 to 1, as 1 indicates that these two documents are very similar to each other, and 0 means that they are completely different. We used the same approach here to compare the topic similarity between our MedHelp and Quora corpus. After deciding the optimal *K*-topics model for each corpus, each topic in each corpus was represented as a vector of words with corresponding prevalence (or the per-topic-per-word probabilities, as the β value). Further, in order to compute the ISC similarity, we only extracted the same words appearing in both corpora. In other words, the *A* and *B* of topic similarity here represented the vector of words and their probabilities in the corresponding topics in each corpus.

Results

Topics generated on MedHelp corpus

With an exploratory approach, we started with testing various *K*-topics models on the MedHelp corpus, where $K = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$. We then narrowed down the optimal topic *K* between 20 to 40. Eventually, we decided that 31 would be the optimal number of *K*-topics model for the MedHelp corpus, based on the model diagnostic metrics and intersubjective qualitative human judgment in the research team. For the details about the decision, please refer to Supplemental Material.

Table 2 displays the 31 topics rendered from the final topic model with $K = 31$, which are ranked based on the topic prevalence across the corpus (*RQ1*). Combining the terms with four different scoring methods (highest probabilities, FREX, Lift and Score) and the representative post(s), we have generated a label for with each topic separately following the common practices in previous studies[36,37].[3] And later we compared and validated all 31 topics together and categoried them into seven groups. The full list of terms for each topic on these four scoring methods could be reviewed in the Supplemental Material.

We can find the top three topics in the MedHelp corpus is Topic 25 *Self-monitor for possible COVID-19 symptoms,* Topic 16 *Face shield/mask for protection*, and Topic 7 *Viral transmission*. We summarized the 31 topics into seven categories, with two topics in diagnosis, six topics in protection and prevention, six topics in pathogenic mechanisms, seven topics in vaccination and vaccine reaction, two topics in public health directives, four topics in social support, and four topics in treatment.

Topics generated on Quora corpus

Adapting the same approach as above, we also started with 11 models whose topics stretched from 10 to 100, and then we tested out a more granular range of models whose *K* lies between 20 to 40. Similarly, we also found that 31-topics model is the best one for Quora corpus based on the model diagnostic metrics and human judgment. The details about the decision could be reviewed in the Supplemental Material as well.

Table 2. The categories and labels of the 31 topics of MedHelp corpus*.

Category	Topics	Labels	Proportion
Diagnosis	**25**	**Self-monitor for possible COVID-19 symptoms**	0.0805
	9	**COVID-19 testing**	0.0517
Protection and prevention	**16**	**Face shield/mask for protection**	0.0588
	20	**Close contact and potential COVID-19 exposure**	0.0433
	10	**Infection prevention and control for COVID-19**	0.0378
	27	Self-protection during the pandemic	0.0295
	3	Against unproved COVID-19 claims	0.0277
	24	Individual response to COVID-19	0.0173
Pathogenic mechanisms	**7**	**Virus transmission**	0.0532
	4	**COVID-19 susceptibility**	0.0457
	28	**Viral pathogenesis of COVID-19**	0.0371
	19	**Wildlife markets and origin of coronavirus**	0.0257
	30	**Concerns of COVID-19 variants**	0.0252
	26	COVID-19 risk factor(s)	0.0222
Vaccination and vaccine reaction	**21**	**Possible side effects after vaccination**	0.0486
	13	**Vaccine skeptics**	0.0448
	11	**Vaccination willingness**	0.0225
	5	Immune response after vaccination	0.0209
	23	**Effectiveness of vaccines**	0.0160
	18	**AstraZeneca vaccine and side effects**	0.0145
	31	**Vaccine hesitancy**	0.0078
Public health directives	**29**	**Lockdown and reopening policy**	0.0361
	6	**Travel restrictions**	0.0191
Social support	**12**	**COVID-19 health anxiety**	0.0355
	22	Caring family during COVID-19 pandemic	0.0307
	17	Quarantine and home care	0.0301
	14	Social support and mental help	0.0245
Treatment	2	Investigational drugs for coronavirus	0.0257
	1	**Coronavirus treatment**	0.0254
	8	Therapeutic management of hospitalized patients	0.0227
	15	**Complementary and alternative medicine**	0.0201

Note: *The highly similar topics (based on the heatmap Figure 4) are highlighted in the table.

Table 3 exhibits all the topics generated from the 31-topics model in Quora corpus, while being ranked based on the topic prevalence (*RQ2*). Similarly, we have labelled these 31 topics of Quora corpus depending on the scoring metrics and human judgement, same as previous section. All the top 20 terms based on different scoring methods corresponding to each topic could be accessed in the Supplemental Material.

The topic modeling discovers the top three topics among Quora corpus are: Topic 8 *Immune and antibody response to vaccination*, Topic 2 *Common reaction to vaccination*, and Topic 4 *Comparing the differences among COVID-19 vaccines*. As Table 3 shows, we further classified the 31 topics into six categories. There are eight topics in vaccine reaction, seven topics in vaccine development

Table 3. The categories and labels of the 31 topics of Quora corpus*.

Category	Topics	Labels	Proportion
Vaccine reaction	**8**	**Immune and antibody responses to vaccination**	0.0698
	2	**Common reaction to vaccination**	0.0552
	24	**Risk comparison for different vaccines**	0.0412
	11	Vaccination perception among vulnerable populations	0.0407
	13	**Historical vaccine safety concerns**	0.0396
	3	Effectiveness of vaccines	0.0393
	31	**Vaccine hesitancy**	0.0332
	30	Vaccine magnetic powers	0.0200
Vaccine development and distribution	**14**	**Comparing the differences among COVID-19 vaccines**	0.0437
	18	**Vaccine research and development**	0.0431
	5	Vaccine manufacturing	0.0408
	4	**Global immunization coverage**	0.0357
	22	Ethical considerations for COVID-19 clinical trials	0.0320
	7	Vaccine distribution	0.0285
	1	The association between vaccine and Luciferase	0.0167
Pathogenic mechanisms	10	Coronavirus lab-leak hypothesis	0.0400
	29	**Concerns of COVID-19 variants**	0.0365
	12	The origin of coronavirus	0.0349
	16	Comparing COVID-19 with other infectious diseases	0.0249
	28	**Virus transmission**	0.0197
	26	**Coronavirus incubation period**	0.0196
Politicization and polarization	17	Attacks on Fauci after email trove released[4]	0.0285
	6	**False and misleading claims toward COVID-19 vaccines**	0.0284
	9	The Trump administration's response to coronavirus	0.0262
	21	**Controversy over mask mandates**	0.0245
	25	Legal debate of vaccine mandates	0.0099
	20	Critics of anti-vaxxers	0.0093
Treatment and control of COVID 19	**27**	**Vaccination rate in controlling of pandemic**	0.0386
	23	**Severity of the coronavirus**	0.0312
	15	**Early treatment for coronavirus**	0.0184
COVID-19 Employee benefits	19	Employer-sponsored health and welfare benefit for employees	0.0297

Note: *The highly similar topics (based on the heatmap Figure 4) are highlighted in the table.

and distribution, six topics in pathogenic mechanism*s*, six topics in politicization and polarization, three topics in treatment and the control of COVID-19, two topics in conspiracy theories, and one standalone topic in *COVID-19* employee benefits.

Topic similarities on MedHelp and Quora

The topic modeling process renders two essential probabilities across the corpus. The first one is θ, which could be denoted as $\theta = \frac{P(Z|X)}{P(X)}$, representing the probability that any given word in the

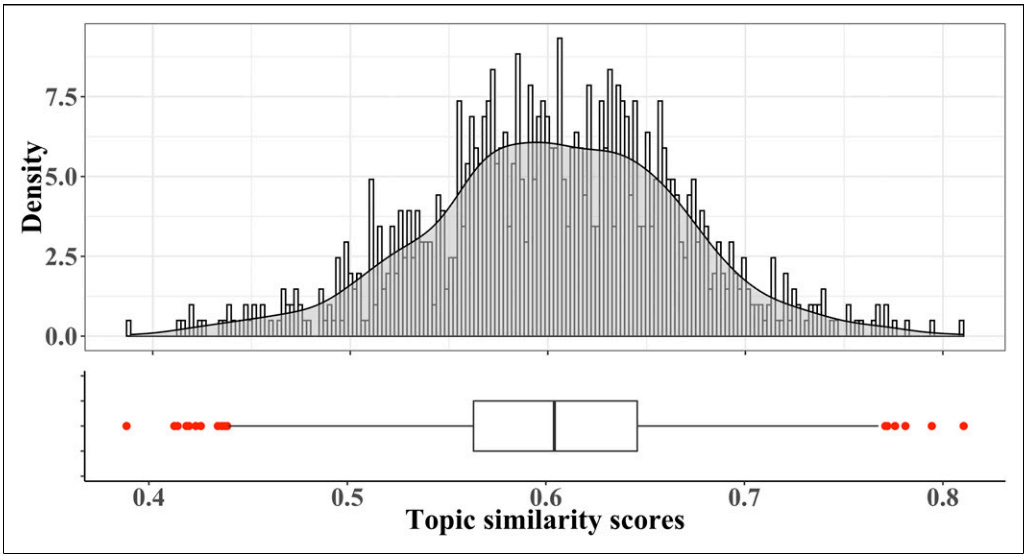

Figure 3. The distribution of topic ISC similarity scores of the 31*31 matrix.

document X that is drawn from topic Z, or the document-topic distribution (the topic proportion on x-axis is Figures 3 and 4). The second one is β, which could be denoted as $\beta = \frac{P(W|Z)}{P(Z)}$, as the topic assignment over all documents Z that contains the specific word W (or the topic-word distribution). In other words, each generated topic is essentially defined by its corresponding probabilities (β) distribution over the vocabulary. In our final models, the MedHelp corpus has 3515 terms and Quora corpus has 6784 terms. To compare the topic similarity, we only extracted the same words that appear in both corpora, which are 3098 terms. Further, we computed the ISC similarities between each vector/topic, which renders a *31*31* matrix. Hereinafter, we are referring the topics of MedHelp corpus as M_n and the topics of Quora corpus as Q_n (for instance, first topic in the MedHelp corpus as M_1 and first topic in the Quora corpus is referred as Q_1). These 961 ISC similarities we computed between MedHelp and Quora topics follow a normal distribution as presented in Figure 3.[5]

Given the ISC similarity scores follow a normal distribution, we set up the arbitrary threshold based on the $mean_{ISC\,similarity} \pm 1.5{*}sd_{ISC\,similarity}$. Two topics are considered as *very similar* when the ISC similarity (M_n, Q_n) is higher than 0.700 (which equal to $mean_{ISC\,similarity} + 1.5{*}sd_{ISC\,similarity}$). And we define two topics are *similar* when the ISC similarity is between 0.603 ($mean_{ISC\,similarity}$) and 0.700. If the ISC similarity of two topics is smaller than 0.603 but larger than 0.506 (which equal to $mean_{ISC\,similarity} - 1.5{*}sd_{ISC\,similarity}$), they are *less similar*. And if the ISC similarity of two topics is smaller than 0.506, they belong to the group of *least similar*. We also used *seriation* R package[38] to reorder the ISC similarity of each pair of topics. And Figure 4 presents the visualization of heatmap after reordering (*RQ3*).

An objective of this study is to explore the similar topics and distinctive topics between MedHelp and Quora. Based on the result of the heatmap, M_1 *Coronavirus treatment*, M_4 *COVID-19 susceptibility*, M_6 *Travel restrictions*, M_7 *Virus transmission*, M_9 *COVID-19 testing*, M_{10} *Infection prevention and control for COVID-19*, M_{11} *Vaccination willingness*, M_{12} *COVID-19 health anxiety*,

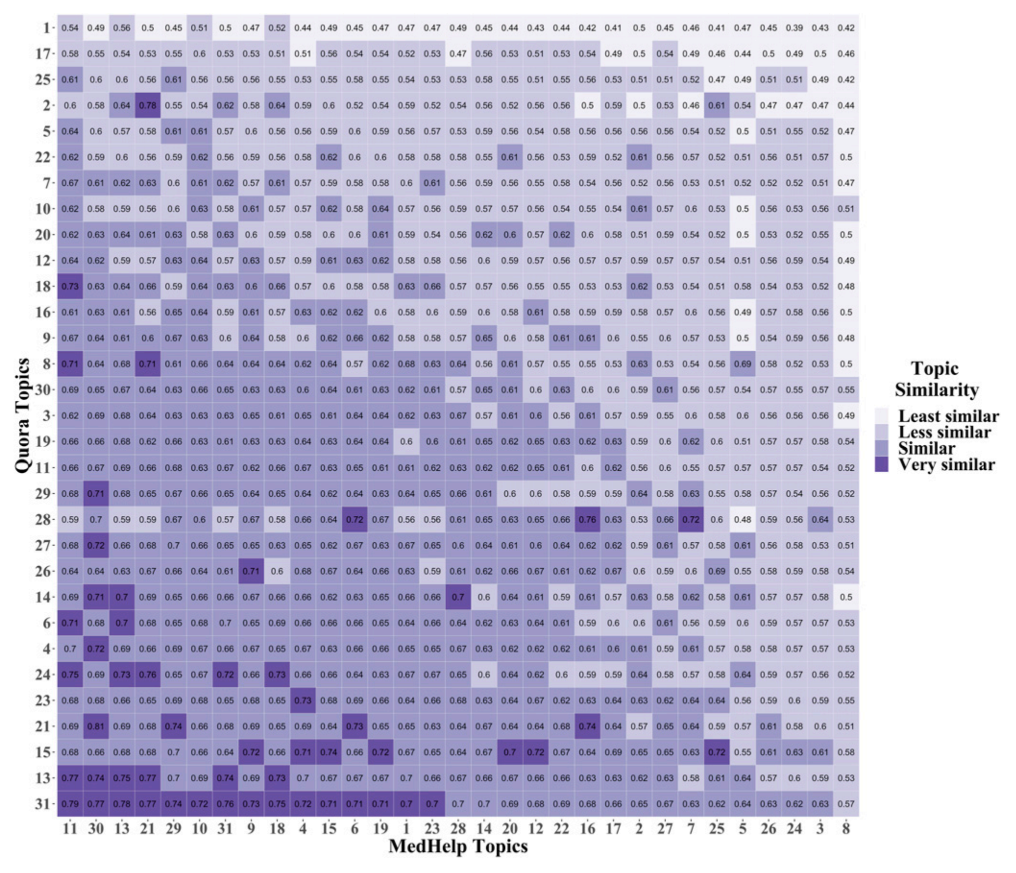

Figure 4. The heatmap of topic ISC similarity scores between MedHelp and Quora topics.

M_{13} *Vaccine skeptics*, M_{15} *Complementary and alternative medicine*, M_{16} *F. shield and mask for protection*, M_{18} *AstraZeneca vaccine and side effects*, M_{19} *Wildlife markets and origin of Coronavirus*, M_{20} *Close contact and potential COVID-19 exposure*, M_{21} *Possible side effects after vaccination*, M_{23} *Effectiveness of vaccine*, M_{25} *Self-monitor for possible covid symptoms*, M_{28} *Viral pathogenesis of COVID-19*, M_{29} *Lockdown and reopening policy*, M_{30} *Concerns of COVID-19 variants*, and M_{31} *Vaccine hesitancy* have a high similar degree with the topics generated from Quora corpus. We can see six out of seven topics in the vaccination and vaccine reaction category, five out of six topics in the pathogenic mechanisms category, three out of six topics in the protection and prevention category, all two topics under the diagnosis category and all two topics under the public health directives category, have a strong and wide similarity degree with topics generated in the Quora corpus. These highly similar topics are highlighted in Tables 2 and 3.

Additionally, we also find Q_2 *Common reaction to vaccination*, Q_4 *Global immunization coverage*, Q_6 *False and misleading claims toward COVID-19 vaccines*, Q_7 *Vaccine distribution*, Q_8 *Immune and antibody responses to vaccination*, Q_{13} *Historical vaccine safety concerns*, Q_{14} *Comparing the differences among COVID-19 vaccines*, Q_{15} *Early treatment for Coronavirus*, Q_{18} *Vaccine research and development*, Q_{21} *Controversy over mask mandates*, Q_{23} *Severity of the*

Coronavirus, Q_{24} *Risk comparison for different vaccines*, Q_{26} *Coronavirus incubation period*, Q_{27} *Vaccination rate in controlling of pandemic*, Q_{28} *Virus transmission*, Q_{29} *Concerns of COVID-19 variants*, and Q_{31} *Vaccine hesitancy* hold strong links with the topics generated from MedHelp corpus. Meanwhile, we can also see five out of seven topics under the vaccine reaction category, three out of seven topics of vaccine development and distribution category, three out of six topics of pathogenic mechanisms category, and all three topics of the treatment and control of COVID-19 category, holds strong links to the MedHelp topics.

Part of the aim of RQ3 is to identify the relatively distinctive topics in each corpus. We find M_8 *Therapeutic management of hospitalized patients* is relatively unique topics in MedHelp since its similarity degrees with other topics are all measured as least similar or less similar. Among the topics generated from Quora, Q_1 *The association between vaccine and Luciferase*, and Q_{17} *Attacks on Fauci after email trove released* are the relatively unique topics.

Discussion

Principle findings

For the first research question, most of the users on MedHelp expressed their confusion and concerns on diagnosis, protection and prevention, coronavirus pathogenic mechanisms, vaccination and vaccine reaction, public health directives, social supports, and treatment. Meanwhile, many users on Quora are concerned about information on vaccine reaction, vaccine development and distribution, pathogenic mechanisms, politicization and polarization, treatment and control of COVID 19, COVID-19 employee benefits. These findings provide a corresponding answer for RQ2. For the third research question, users on MedHelp and Quora expressed same concerns on vaccination, including possible vaccine reaction, safety and effectiveness of possible vaccine, vaccine development and distribution. A considerable number of users on MedHelp and Quora also cared for basic elements of the disease including the onset symptoms, transmission routes, preventive measures, treatment and control of COVID-19.

There are some differences of users' information needs on COVID-19 between the OHC and the Q&A forum. One major difference is that the proportion of posts reported being worried for themselves and their close ones on MedHelp was higher than Quora. This finding reveals that users on MedHelp have more information needs on psychological health, which accords with earlier observation reported by Chen et al.[39] The reason for the topic disparities might be the anonymity feature of the OHC. It is easier to build a strong relationship within OHC, which led users to describe their concerns and conditions more carefully, including more details and personal feelings on Medhelp. Previous evidence suggests that one of the main reasons people participate in an OHC is to seek and obtain various types of social support,[40] and social bonds grow stronger during times of uncertainty and crisis.[41] Thus, while individual experiences during COVID-19 are nuanced, OHCs such as Medhelp could serve to facilitate connections and provide opportunities for new modes of interaction and meaningful relationship-building during the pandemic.

Quora is a Q&A platform where diverse perspectives and voices are shared and heard in a culture of knowledge sharing.[42] We found that the current dominant users of Quora well represent the users of the online knowledge community, as Quora users hold more profound perspectives compared with users on Medhelp. For instance, the discussion around vaccination on Medhelp mainly focuses on individual-level risk/benefit balance, while vaccination discussions on Quora would also surround social ethics such as the ethical considerations of COVID-19 clinical trials. The results also show that Quora users are more active in discussions around controversial topics, such as anti-vaccine statements and political divides. In that sense, information exchanges on these topics were not only about exchanging helpful knowledge but also imbued with conspiracy beliefs and political judgment.

Practical implications

There are widespread public concerns about vaccine, including reactions following a vaccination, safety and effectiveness of possible vaccine, vaccine development and distribution. Although evidence has demonstrated that vaccinations were currently the best hope society had to contain the pandemic, fear, and confusion still muddled people's confidence in vaccine. Concerns about approvals being rushed, suspicion of the pharmaceutical industry, uncertainty surroundings the vaccine are the widely mentioned reasons for vaccine hesitancy.[43] Public health authorities and pharmaceutical manufacturers must continue to communicate transparently with the public about any potential side effects of COVID-19 vaccines when highlight vaccines' effectiveness. The Food and Drug Administration (FDA) and Centers for Disease Control and Prevention (CDC) in the United States have been thorough and transparent about the safety profiles of each of the vaccines. One example is that in April 2021 the CDC and FDA paused the use of the Johnson and Johnson vaccine and then further investigated rare blood clots that had been reported after vaccination.

Health professionals and communicators must help individuals emotionally cope with the psychological trauma during the COVID-19 pandemic and navigate to a post-crisis new normal. Users on Medhelp have expressed their needs regarding social support during the pandemic. As individuals practice social distancing and quarantine in an effort to help prevent the spread of Coronavirus, they may experience a higher prevalence of loneliness, feelings of isolation, and poor mentor health. Research has demonstrated that having strong social support during times of crisis can help mitigate mental disorders.[44] Public health authorities and communicators should intervene and fill the support gap for public during the pandemic. For instance, healthcare could be delivered remotely through several telehealth modalities and treatment protocols to provide medical assistance and psychotherapy services. Medical staff shall pay attention to the emotional and psychological conditions of patients, encourage them to speak openly about their concerns, and provide support when needed. Government should provide resources and social support services such as disaster financial assistance with food, housing, and bills.

Quora has unique features, such as the real name environment and the upvote feature, making Quora a more proper platform to rationally discuss health issues rather than to emotionally express personal attitudes. Another strength of Q&A websites is that multiple users can answer the same question, offering more than one explanation from different perspectives, which could aid in reconciling varying viewpoints and have more conciliatory conversations. The echo chambers in the online environment can thus be disrupted by providing balanced arguments. We found that some discourse on Quora becomes more polarized on the issue of handling the Coronavirus crisis. Public opinion is deeply divided along the views on urgency of the crisis, responsiveness to government decisions and personal behavioral responses to the COVID-19 pandemic. This finding is consistent with previous observations that partisan gaps exist in views of many aspects of the pandemic such as risk perceptions and responses to pandemics.[45,46] Future research could investigate the effects of politicized and polarized online information on the community vulnerability to COVID-19, as well as strategies to reduce divisions and break away from previous patterns of reflexive partisanship.

Limitations

This paper has a few limitations that could be addressed in the future. This study has taken a post as the unit of analysis to adjust the length issue of online post, which might miss nuances within each answer. The future research could employ bi-term topic modeling or single topic LDA to incorporate more flexibility within each answer or incorporate user-defined seed words for topic-word distribution to better address the domain-specificity problem.[47] Also, we acknowledge the subjective judgement when assigning the topic labels and the possible bias could affect the conclusion. In the future, we can incorporate a generative labeling approach to help with evaluate the topic quality.[48] During the text preprocessing, thought we have followed the common practices in the field including removing stopwords and performing stemming, we should be more cautious the actions as some of them might have limited effects on improving the performance of topic modeling as Schofield et al. argue.[49,50] Besides, this study has regarded all the posts as static. However, in reality we know the temporal trend would be an important element to capture the dynamic of information needs. In the future research, we can scrape the posting time for each answer, and conduct structural topic modeling to examine the fluctuation of the topic perplexity across time. To be noted, Quora does not show the exact date if it was posted more than 1 year. Lastly, when we computed the ISC similarity score between the MedHelp and Quora corpus, we only used the overlapped terms which could ignore the unique linguistic pattern especially in the Quora corpus (since only 3098 out of 6784 terms were included). In the future, we may compute the similarity score based on the sparse matrix with all the terms involved in both corpora, and examine the differences compared to the approach in this study.

Conclusion

Coronavirus was a heated discussion topic on both MedHelp and Quora. In this paper, we examined the COVID-19 related posts on OHC Medhelp and on Q&A platform Quora and identified a variety of information needs of the general public over the course of the pandemic. Moreover, we discovered the disparities of the information needs between the users of these two online platforms. To best of our knowledge, this is the first study to examine the public's information needs comparing two different one-to-many knowledge-based online platforms during the COVID-19 pandemic. This insight is beneficial for tracking and responding to the public's information needs during pandemic. The findings from this study could also provide refined knowledge for researchers or practitioners who aim to provide accurate information assistance and build effective online emergence response programs.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Availability of data and materials

The code for the web scraping is available on https://github.com/ryanyangwang/hij_information_needs.

Ethical statement

All the data collected in the study is readily available to the public and there is no direct interaction with participants during the web scraping process. No personal information on user-level was obtained or stored and there is no possible way to link a record with a particular individual. Therefore, no ethical approval is required in the current study.

ORCID iD

Ryan Yang Wang https://orcid.org/0000-0002-1875-0438

Supplemental Material

Supplemental material for this article is available online.

Notes

1. The top country with the most traffics for both MedHelp and Quora is the United States based on similarweb: https://www.similarweb.com/website/medhelp.org/#traffic; and https://www.similarweb.com/website/quora.com/#traffic.
2. STM it is a generalization/combination of CTM, Dirichlet-Multinomial Regression (DMR) topic model, and Sparse Additive Generative Model (SAGE).[28] One generalization of CTM in the STM is the replacement of the logistic normal prior with a logistic normal linear model. However, within a STM without structural metadata, the linear model is reduced to the same value. Therefore, it is a similar result but not exactly the same as the SAGE and DMR elements are different in the model.
3. Two experts in public health and computational linguistics in the research team evaluated the results of the topic models and identified the labels for each topic (of each corpus). Specifically, one expert started with reading the top 20 terms and representative documents for each topic and suggested the annotation of each topic. Then the other expert validated and complemented the annotation done by the first expert separately while repeating the same process as above. Then both experts met and made an agreement together on the final label. If the experts could not meet the consensus on a specific topic, they went through more terms and more documents for the topic together until the agreement was met.
4. On 1 June 2021, the Washington Post (886 pages) and BuzzFeed News (3200 pages) released the emails of the director of National Institute of Allergy and Infectious Diseases, Dr. Anthony Fauci from January 2020 to June 2020, as part of a Freedom of Information Act request. For more details, please refer to the Washington Post: https://www.washingtonpost.com/politics/interactive/2021/tony-fauci-emails/; and BuzzFeed News via https://www.buzzfeednews.com/article/nataliebettendorf/fauci-emails-covid-response.
5. Although the Shapiro-Wilk normality test shows the p-value is slightly higher than 0.05 ($p = 0.062$), the p-value of Anderson-Darling normality test is slightly lower than 0.05 ($p = 0.049$). The skewness test (skewness = −0.130) finds the data is slightly negative skewed but still within the range of [-0.5, 0.5]. Thus, the ISC similarity scores among MedHelp and Quora topics follow a symmetric and normal distribution. Minimum value is 0.389, maximum value is 0.811, mean value is 0.603, and the standard deviation is 0.065.

References

1. Pew Research Center. 2013. Health Online 2013. https://www.ordinedeimedici.com/documenti/Docs7-cybercondria-PIP-HealthOnline.pdf
2. Li J, Theng YL and Foo S. Predictors of online health information seeking behavior: changes between 2002 and 2012. *Health Informatics J* 2016; 22(4): 804–814. DOI: 10.1177/1460458215595851

3. Xiang YT, Yang Y, Li W, et al. Timely mental health care for the 2019 novel coronavirus outbreak is urgently needed. *Lancet Psychiatry* 2020; 7(3): 228–229. DOI: 10.1016/S2215-0366(20)30046-8
4. Barry JM. Pandemics: avoiding the mistakes of 1918. *Nature* 2009; 459(7245): 324–325. DOI: 10.1038/459324a
5. Garfin DR, Silver RC and Holman EA. The novel coronavirus (COVID-2019) outbreak: amplification of public health consequences by media exposure. *Health Psychol* 2020; 39(5): 355. DOI: 10.1037/hea0000875
6. Merchant RM. Evaluating the potential role of social media in preventive health care. *JAMA* 2020; 323(5): 411–412. DOI: 10.1001/jama.2019.21084
7. Lachlan KA, Spence PR, Lin X, et al. Social media and crisis management: CERC, search strategies, and Twitter content. *Comput Hum Behav* 2016; 54: 647–652. DOI: 10.1016/j.chb.2015.05.027
8. Burzyńska J, Bartosiewicz A and Rękas M. The social life of COVID-19: Early insights from social media monitoring data collected in Poland. *Health Informatics J* 2020; 26(4): 3056–3065. DOI: 10.1177/1460458220962652
9. Chen E, Lerman K and Ferrara E. Tracking social media discourse about the covid-19 pandemic: development of a public coronavirus twitter data set. *JMIR Public Health Surveill* 2020; 6(2): e19273. DOI: 10.2196/19273
10. Tshimula JM, Chikhaoui B and Wang S. COVID-19: detecting depression signals during stay-at-home period. *Health Informatics J* 2022; 28(2): 14604582221094931. DOI:10.1177/1460458222109493
11. Bopape S, Mahlatji MM, Dikotla M, et al. Identifying the information needs of public library and information services users in Limpopo province. *South Afr J Libraries Inf Sci* 2017; 83(1): 1. DOI:10.7553/83-1-1631
12. Nicholas D. *Assessing information needs: tools, techniques and concepts for the internet age*. London, UK: Routledge, 2003.
13. Ormandy P. Defining information need in health–assimilating complex theories derived from information science. *Health Expectations* 2011; 14(1): 92–104. DOI: 10.1111/j.1369-7625.2010.00598.x
14. Zhao W, Lu P, Yu S, et al. Consumer health information needs in China–a case study of depression based on a Social Q&A community. *BMC Med Inform Decis Mak* 2020; 20(3): 1–9. DOI: 10.1186/s12911-020-1124-1
15. Springer S, Menzel LM and Zieger M. Google trends provides a tool to monitor population concerns and information needs during COVID-19 pandemic. *Brain Behav Immun* 2020; 87: 109–110. DOI: 10.1016/j.bbi.2020.04.073
16. Wei J, Huang C, Vosoughi S, et al. What are people asking about COVID-19? A question classification dataset. In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Online, 2020. https://aclanthology.org/2020.nlpcovid19-acl.8
17. Chen AT. Exploring online support spaces: using cluster analysis to examine breast cancer, diabetes and fibromyalgia support groups. *Patient Educ Couns* 2012; 87(2): 250–257. DOI: 10.1016/j.pec.2011.08.017
18. Park H and Park MS. Cancer information-seeking behaviors and information needs among Korean Americans in the online community. *J Community Health* 2014; 39(2): 213–220. DOI: 10.1007/s10900-013-9784-8
19. Liu J, Kong J and Zhang X. Study on differences between patients with physiological and psychological diseases in online health communities: topic analysis and sentiment analysis. *Int J Environ Res Public Health* 2020; 17(5): 1508. DOI: 10.3390/ijerph17051508
20. Oh S. Chapter 3: social Q&A. In: Brusilovsky P and He D (eds) *Social information access: systems and technologies*. 1st ed. New York, NY: Springer, 2018, pp. 75–107.
21. Naderi H, Madani S, Kiani B, et al. Similarity of medical concepts in question and answering of health communities. *Health Informatics J* 2020; 26(2): 1443–1454. DOI: 10.1177/1460458219881333

22. Zhao Y and Zhang J. Consumer health information seeking in social media: a literature review. *Health Info Libr J* 2017; 34(4): 268–283. DOI: 10.1111/hir.12192
23. Oh S, Zhang Y and Park MS. Cancer information seeking in social question and answer services: identifying health-related topics in cancer questions on Yahoo! Answers. *Inf Res* 2016; 21(3), http://informationr.net/ir/21-3/paper718.html
24. Rizvi RF, Wang Y, Nguyen T, et al. Analyzing social media data to understand consumer information needs on dietary supplements. *Stud Health Technol Inform* 2019; 264: 323. DOI: 10.3233/SHTI190236
25. Yang CC and Tang X. Estimating user influence in the MedHelp social network. *IEEE Intell Syst* 2012; 27(05): 44–50. DOI: 10.1109/MIS.2010.113
26. Quora. Reach 300+ million monthly unique visitors on Quora. https://business.quora.com/Reach-300-million-monthly-unique-visitors-on-Quora (Accessed October 23, 2021).
27. Selenium. WebDriver. https://www.selenium.dev/documentation/webdriver/ (Accessed October 16, 2021).
28. MedHelp. Coronavirus Community. https://www.medhelp.org/forums/Coronavirus/show/2203 (Accessed October 16, 2021).
29. Quora. COVID-19. 2019-2020. https://www.quora.com/topic/COVID-19 (Accessed October 16, 2021).
30. Slemon A, McAuliffe C, Goodyear T, et al. Reddit users' experiences of suicidal thoughts during the COVID-19 pandemic: a qualitative analysis of r/Covid19_support posts. *Front Public Health* 2021; 9: 693153. DOI: 10.3389/fpubh.2021.693153
31. Roberts ME, Stewart BM and Tingley D. stm: an R package for structural topic models. *J Stat Softw* 2019; 91(1): 1–40. DOI: 10.18637/jss.v091.i02
32. Gomaa WH and Fahmy AA. A survey of text similarity approaches. *Int J Comput Appl* 2013; 68(13): 13–18. DOI: 10.5120/11638-7118
33. Aggarwal CC, Hinneburg A and Keim DA. On the surprising behavior of distance metrics in high dimensional space. In: Van den Bussche J and Vianu V (eds) *Database theory—ICDT 2001*. London, UK: Springer, 2001, pp. 420–434.
34. Zhu S, Liu L and Wang Y. *Information retrieval using Hellinger distance and sqrt-cos similarity*. In: 2012 7th international conference on computer science & education (ICCSE), Melbourne, Australia, 2012, pp. 925–929. IEEE. DOI: 10.1109/ICCSE.2012.6295217
35. Sohangir S and Wang D. Improved sqrt-cosine similarity measurement. *J Big Data* 2017; 4(1): 1–3. DOI: 10.1186/s40537-017-0083-6
36. Moßburger L, Wende F, Brinkmann K, et al. Exploring online depression forums via text mining: a comparison of Reddit and a curated online forum. In: Proceedings of the fifth social media mining for health applications workshop & shared task, Barcelona, Spain (online), 2021, pp. 70–81. https://aclanthology.org/2020.smm4h-1.11
37. Gao S, Pandya S, Agarwal S, et al. Topic Modeling for Maternal Health Using Reddit. In: Proceedings of the 12th international workshop on health text mining and information analysis, Online, 2021, pp. 69–76. https://aclanthology.org/2021.louhi-1.8
38. Hahsler M, Hornik K and Buchta C. Getting things in order: an introduction to the R package seriation. *J Stat Softw* 2008; 25(3): 1–3. DOI: 10.18637/jss.v025.i03
39. Chen Y, Dong T, Ban Q, et al. What concerns consumers about hypertension? A comparison between the online health community and the Q&A forum. *Int J Comput Intell Syst* 2021; 14(1): 734–743. DOI: 10.2991/ijcis.d.210203.002
40. Kim E, Han JY, Moon TJ, et al. The process and effect of supportive message expression and reception in online breast cancer support groups. *Psychooncology* 2012; 21(5): 531–540. DOI: 10.1002/pon.1942
41. Haslam SA, Reicher SD and Platow MJ. *The new psychology of leadership: Identity, influence and power*. New York, NY: Routledge, 2020.

42. Quora. Why Quora exist. https://www.quora.com/about (Accessed October 23, 2021).
43. Lane S, MacDonald NE, Marti M, et al. Vaccine hesitancy around the globe: analysis of three years of WHO/UNICEF Joint Reporting Form data-2015–2017. *Vaccine* 2018; 36(26): 3861–3867. DOI: 10.1016/j.vaccine.2018.03.063
44. Saltzman LY, Hansel TC and Bordnick PS. Loneliness, isolation, and social support factors in post-COVID-19 mental health. *Psychol Trauma* 2020;12(S1):S55. DOI: 10.1037/tra0000703
45. Allcott H, Boxell L, Conway J, et al. Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *J Public Econ* 2020; 191: 104254. DOI: 10.1016/j.jpubeco.2020.104254
46. Clinton J, Cohen J, Lapinski J, et al. Partisan pandemic: how partisanship and public health concerns affect individuals' social mobility during COVID-19. *Sci Adv* 2021; 7(2): eabd7204. DOI: 10.1126/sciadv.abd7204
47. Tushev M, Ebrahimi F and Mahmoud A. Domain-specific analysis of mobile app reviews using keyword-assisted topic models. In: Proceedings the 44th international conference on software engineering, Pittsburgh, US, 2022, pp. 762–773. DOI: 10.1145/3510003.3510201
48. Alokaili A, Aletras N and Stevenson M. Automatic generation of topic labels. In: Proceedings the 43rd international ACM SIGIR conference on research development information retrieval, Online, 2020, pp. 1965–1968. DOI: 10.1145/3397271.3401185
49. Schofield A and Mimno D. Comparing apples to apple: the effects of stemmers on topic models. *Trans the Assoc Comput Linguistics* 2016; 4: 287–300. DOI: 10.1162/tacl_a_00099
50. Schofield A, Magnusson M and Mimno D. *Pulling out the stops: rethinking stopword removal for topic models*. In: Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 2, Valencia, Spain, 2017, pp. 432–436. https://aclanthology.org/E17-2069