Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Computers in Biology and Medicine 132 (2021) 104336

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

Twitter-based analysis reveals differential COVID-19 concerns across areas with socioeconomic disparities

Yihua Su [a,1], Aarthi Venkat [b,1], Yadush Yadav [a], Lisa B. Puglisi [c,d], Samah J. Fodeh [a,b,e,*]

[a] Health Informatics Program, Yale School of Public Health, 60 College St, New Haven, CT, 06510, USA
[b] Computational Biology and Bioinformatics Program, Yale University, 300 George Street, Suite 501, New Haven, CT, 06511, USA
[c] SEICHE Center for Health and Justice, Yale School of Medicine, 333 Cedar St, New Haven, CT, 06510, USA
[d] Pain Research, Informatics, Multimorbidities and Education Center, VA Connecticut Healthcare System, 950 Campbell Avenue, West Haven, CT, 06516, USA
[e] Department of Emergency Medicine, Yale School of Medicine, 333 Cedar St, New Haven, CT, 06510, USA

ARTICLE INFO

Keywords:
COVID-19
Twitter
Social media
Socioeconomic status
Topic modeling

ABSTRACT

Objective: We sought to understand spatial-temporal factors and socioeconomic disparities that shaped U.S. residents' response to COVID-19 as it emerged.
Methods: We mined coronavirus-related tweets from January 23rd to March 25th, 2020. We classified tweets by the socioeconomic status of the county from which they originated with the Area Deprivation Index (ADI). We applied topic modeling to identify and monitor topics of concern over time. We investigated how topics varied by ADI and between hotspots and non-hotspots.
Results: We identified 45 topics in 269,556 unique tweets. Topics shifted from early-outbreak-related content in January, to the presidential election and governmental response in February, to lifestyle impacts in March. High-resourced areas (low ADI) were concerned with stocks and social distancing, while under-resourced areas shared negative expression and discussion of the CARES Act relief package. These differences were consistent within hotspots, with increased discussion regarding employment in high ADI hotspots.
Discussion: Topic modeling captures major concerns on Twitter in the early months of COVID-19. Our study extends previous Twitter-based research as it assesses how topics differ based on a marker of socioeconomic status. Comparisons between low and high-resourced areas indicate more focus on personal economic hardship in less-resourced communities and less focus on general public health messaging.
Conclusion: Real-time social media analysis of community-based pandemic responses can uncover differential conversations correlating to local impact and income, education, and housing disparities. In future public health crises, such insights can inform messaging campaigns, which should partly focus on the interests of those most disproportionately impacted.

1. Introduction

Early in the course of the COVID-19 pandemic, with no specific treatment for the disease available and fears of the burden of illness overwhelming health systems, the primary public health focus was on disease mitigation strategies [1–3] – and it still is almost a year later. New concepts were introduced to the general public, such as social distancing and recommendations for routine masking. These mitigation efforts along with others, including travel bans, shelter-in-place orders, and school closures, were anticipated to negatively affect many sectors of the United States (U.S.) economy, and they have drastically changed the quotidian lives of most Americans. Given marked community-level socioeconomic disparities and segregation in the U.S. that predated COVID-19, these measures were likely to have disparate uptake by and impact on Americans depending on where they live [4].

With the expansive geography of the U.S. and modern-day travel patterns, the disease initially was largely localized in a few cities, and these so-called "hotspots" were a primary focus of much of the initial media coverage [5]. However, as expected, other COVID-19 hotspots with large marginalized populations later emerged [6,7]. This brought to the forefront the need to understand differential reactions to the crisis as a tool for shaping public health communication and allocation of

* Corresponding author. 300 George Street, PO Box 208009, New Haven, CT, 06520, USA.
E-mail address: samah.fodeh@yale.edu (S.J. Fodeh).
[1] Co-first authors, contributed equally to this work.

https://doi.org/10.1016/j.compbiomed.2021.104336
Received 23 November 2020; Received in revised form 8 March 2021; Accepted 10 March 2021
Available online 13 March 2021

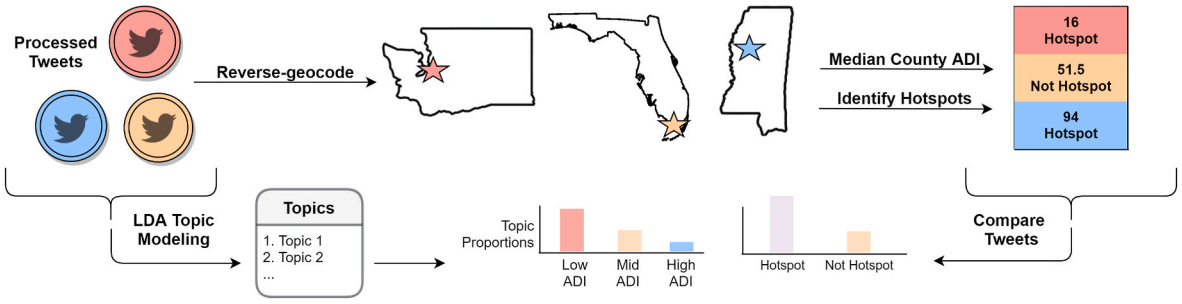

Fig. 1. Data integration and analysis workflow.

health resources.

Social media has been a prominent venue for personal and public health communication, both in previous public health crises and now with COVID-19. Twitter, in particular, has the advantage over some other social media platforms of providing brief, real-time content availability with access to networks of similar discussions through hashtags. Twitter has been used to assess mitigation strategies such as social distancing [8,9] and estimate mobility dynamics within and across states [10]. However, Twitter has not, to our knowledge, been used as a tool to identify trends in public responses to a health crisis at the local level, while factoring in socioeconomic status. Understanding the public responses and reactions at the initial stage of the pandemic across areas with socioeconomic disparities better inform future public health guidelines and communication under similar circumstances.

In this study, we sought to leverage a novel approach that utilizes Twitter to understand how social media analysis can provide insight on local level concerns that can guide future public health pandemic messaging. Specifically, we investigated two hypotheses that: 1. there are differential concerns across less-resourced areas (low ADI) and high-resourced areas (high ADI) and, 2. there exist differential concerns across hotspots and non-hotspots.

In the following section, we provide a brief review of the related literature, and in Material and Methods, we describe the Twitter data and implemented methods used for analysis. In the Results section, we present our findings and discuss and comment on them in the Discussion section. We also report the limitations of the study in the Limitation subsection and finally conclude the paper in the Conclusion section.

2. Related work

To provide greater context for understanding our use of Twitter in this study, we first provide a general and brief review of how natural language processing and analytics of Twitter data have been used as research and public health tools to characterize, contextualize and monitor health conditions. Pre-COVID-19, social media research in the context of health was primarily focused on examining the patient experience [13–17]. Comments and reviews on Twitter were used to measure healthcare quality [15] and monitor patient health status along with sentiment level [17]. It has also been useful in understanding social networks, public health messaging, and forecasting spread [19–22]. Twitter played an important role in Ebola outbreak surveillance by contributing to disease surveillance efforts – detecting an epidemic nearly a week before its first case [20]. Influenza infection rate [21] and Zika Virus case number [22] predictions, learned from the tweet count pattern of disease-related tweets, have also proven successful.

During COVID-19, Twitter has been used to capture self-reported symptoms of COVID-19 [23] and explore fake news and rumors related to the pandemic [24]. Many studies have explored the utility of using advanced data analytics such as neural networks to study the spread and impact of COVID-19 [25]. Different types of data were utilized in these studies, including; medical image data harnessed for early detection of COVID-19 [26], mortality and recovery rates leveraged to measure the security levels of the pandemic [27], and mobility data of cellphone users for monitoring impacts on the spread of COVID-19 [28,29]. Twitter data has also been used to learn more about COVID-19 spread and impact. It has been used to assess mitigation strategies such as social distancing [18,19], capture self-reported symptoms of COVID-19 [20], and identify differential psychological impacts of lockdowns using hashtags [30].

3. Materials and methods

3.1. Twitter dataset

The dataset we used for this analysis is composed of Twitter entries (tweets) in English posted by users in the United States from January 23rd to March 25th, 2020. We mined the tweets with Twitter's standard search API, which returns a sampling of relevant tweets matching a specific query [31]. This search service is not meant to be an exhaustive source of tweets, and is instead optimized for relevance to the query. We queried for keywords 'coronavirus', 'corona virus', 'corona', 'covid', 'covid-19', 'covid 19', and 'covid19'. For each tweet, the Twitter standard search API provides detailed tweet attributes, including unique de-identified user ID, time and text of the tweet, and four geographic coordinates (latitude and longitude) delineating the bounding box [32] from which the tweet was posted. For privacy reasons, Twitter does not provide the exact location from which tweets were posted. Fig. 1 demonstrates the overall workflow of the analysis given these tweet attributes, which will be further detailed in the following sections.

3.2. Preprocessing of tweets

We pre-processed the tweets following standard data cleaning practices [33] through the removal of punctuation marks, numbers, emojis, URLs, stop words, and end of line characters. We then shortened the remaining words to the root using the stemmer package provided by the NLTK toolkit [34]. We removed tweets that were with missing or invalid data such as those without a month or date of entry, valid user ID entry, or valid stemmed tweet text. Finally, we filtered out tweets containing only words that occurred in less than 20 documents or more than 50% of all documents (of which only "coronavirus" was excluded) in order to achieve better topic models. This is a common approach [35,36], used to avoid spurious associations by excluding words based on their frequency distribution.

3.3. Reverse geocodes of tweets

We employed GeoPy [37] to reverse geocode the coordinates and output the county and state names of each tweet. As the bounding box provides enough information to confidently geotag the tweet at the county resolution, we used the midpoint of the rectangle of latitude and longitude coordinates of each tweet as the effective location. This location was then linked to a five-digit FIPS code, a code designed to uniquely identify counties and states in the U.S., to determine the

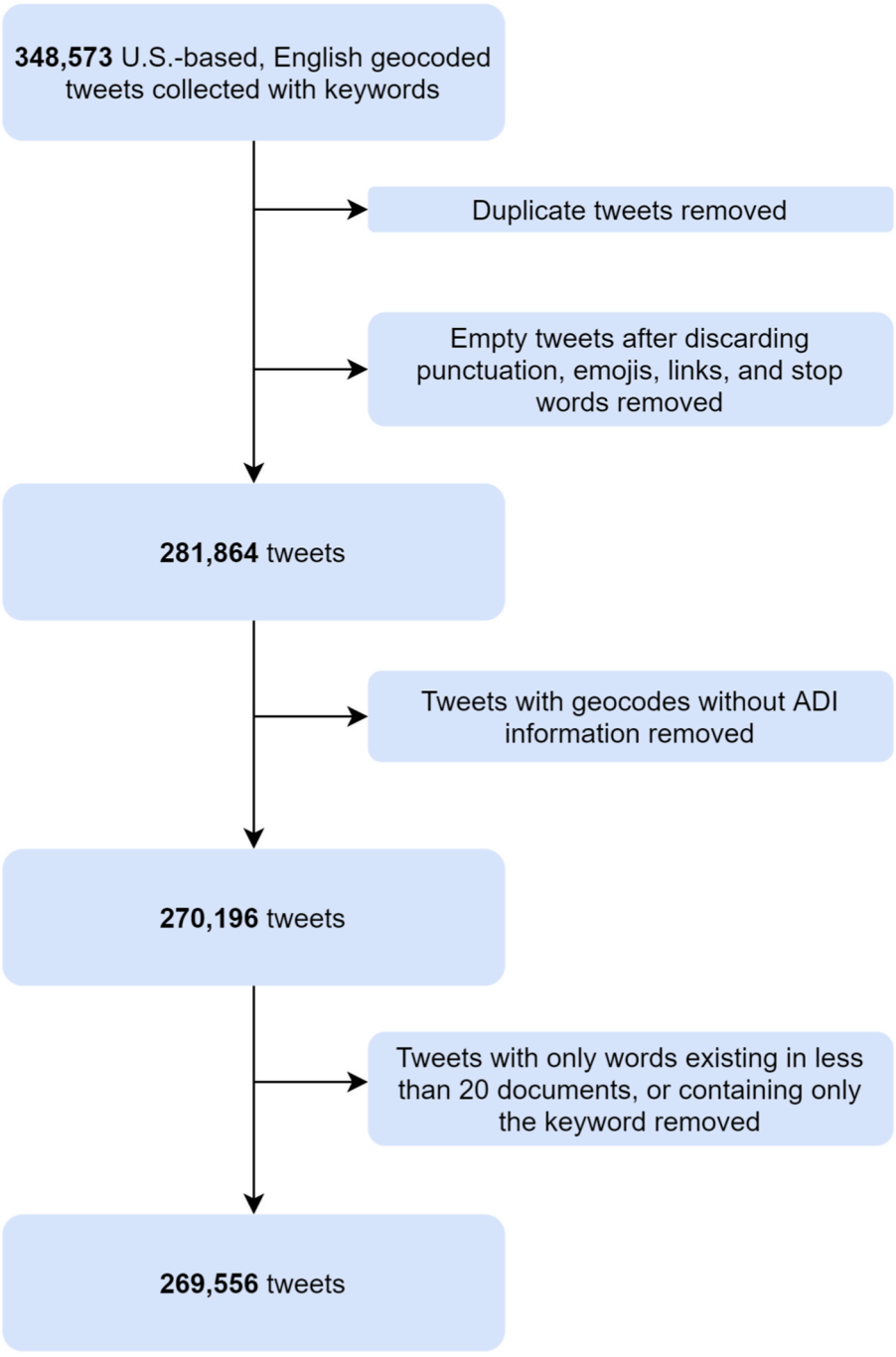

Fig. 2. Object process diagram of tweet pre-processing.

location of tweets at the county level. We followed a similar approach in our previous work [9] to map tweets to the county level.

3.4. Area Deprivation Index (ADI) designation

We leveraged ADI from The Neighborhood Atlas [38], a location-based socioeconomic index at the census-block-group level, which incorporates income, education, employment, and housing data and has been used to inform health delivery and policy. ADI scores range from 0 to 100, where 0 corresponds to low deprivation and 100 corresponds to high deprivation. We mapped the location of each tweet, derived from the reverse geocoding tweets process, to the median ADI score of all the census block groups within the county using its FIPS code. Counties were considered "low", "mid", or "high" ADI based on the ADI distribution of the unique counties represented in the dataset. Low ADI designation was assigned to counties from the lowest quintile of the ADI distribution of represented counties, and high ADI designation was assigned to counties from the highest quintile of the distribution as has been done with other studies using ADI [39,40].

3.5. Hotspot identification

We defined hotspots in January and February as areas with any cases of COVID-19 because there were few U.S. cases in these months and they were concentrated (as published by the New York Times [41]). For analyzing hotspots in March, we leveraged the curated resource The U.S. COVID-19 Atlas [42], defining a tweet as from a hotspot if the county was listed among the published population-adjusted hotspots.

3.6. Topic modeling

We used the Latent Dirichlet Allocation (LDA) approach [43] for topic modeling. LDA is an unsupervised approach and has shown to be successful at modeling topics in tweets [44]. We leveraged LDA from the MALLET package [45] and "gensim" package in Python to detect topics from COVID-related tweets. To determine the optimal number of topics, we compared topics by their coherence scores, which act as a proxy for interpretability by measuring the degree of semantic similarity between top words in the topic [46]. We used the topic-word distribution to annotate topics. We first ranked words of a topic and then assigned the underlying theme.

3.7. Spatiotemporal analysis

We leveraged the document-topic probability distribution for this analysis. We compared topic prevalence over time, across low and high ADI areas, between hotspots and non-hotspots areas, and within hotspots between low ADI and high ADI areas.

3.7.1. Temporal analysis of topic prevalence

To understand how the public reactions to COVID-19 varied temporally, we averaged the topic distributions of all tweets for each month. We then compared the average scores of all topics over time. For selected topics, we plotted out the daily topic dynamic to demonstrate how the topic distribution changed.

3.7.2. Spatial analysis of topic prevalence

We anticipated that the topic differences across areas with differing ADIs would be skewed, thus we used the log of odds ratio (log odds ratio), a common approach to transform skewed data to a normal distribution [47], to compare the topic differences across area groups. To compare the dominant topics in counties of low versus high ADI designation, we computed the log odds ratios of dominant topics in both groups. We first identified the dominant topics – the topics with the highest probability – for all tweets, then we calculated the log odds ratio of dominant topics among both groups to achieve a fair comparison. The log odds ratio of a topic can be interpreted as the probability of dominance of that topic in one group over another.

The odds that any topic T dominates in a group G are calculated as:

$$odd(\mathrm{T},\mathrm{G}) = \frac{number\ of\ tweets\ that\ topic\ T\ is\ dominant\ in\ group\ G}{total\ tweets\ in\ group\ G}$$

The log odds ratio of any topic T between two groups G_0, G_1 is calculated as:

$$log\ odds\ ratio\ (T,\ G_0,\ G_1) = log(\frac{odd(T,G_0)}{odd(T,G_1)})$$

We used the same approach to compare topic prevalence between hotspots and non-hotspots. All of the calculations above were done in Python, using the packages "NumPy" and "math".

3.8. Statistical validation

We implemented the chi-squared test and independent *t*-test to assess the differences in discussed topics across geographically grouped tweets. More specifically, the chi-squared test was used to validate the hypotheses stated in the Introduction Section that 1. there are differential concerns across less-resourced areas (low ADI) and high-resourced areas (high ADI) and, 2. there exist differential concerns across hotspots and non-hotspots.

The chi-squared test determines whether there were statistically significant differences between the expected dominant topic frequencies and observed dominant topic frequencies across the ADI groups and hotspot groups. And according to related researches, we acknowledged the nature of Twitter data might be imbalanced [48,49] and further leveraged Welch's unequal variances *t*-test, which is more robust than Student's t-test for skewed distributions and unequal sample sizes [50], to identify the topics that have significant differences between the groups. Formally, the *t*-test determines whether there was a difference between the means of the dominant topic probabilities in the low and high ADI groups. All of the statistical validations above were conducted through SPSS.

Table 1
Characteristics of Dataset. Summary statistics of Twitter dataset in terms of user, geographic, and socioeconomic distribution.

Tweet Characteristics (n = 269,556)	
Southern States	36.65% (n = 98,792)
Western States	28.10% (n = 75,745)
Northeastern States and DC	20.42% (n = 55,043)
Midwestern States	14.64% (n = 39,462)
Puerto Rico	0.19% (n = 514)
Low ADI (3-43)	50.07% (n = 134,967)
Mid ADI (43.5–77)	45.65% (n = 123,052)
High ADI (77.5–98)	4.29% (n = 11,537)
Mean Tweet Count per User	2.25 tweets
Median Tweet Count per User	1 tweet
Max Tweet Count per User	456 tweets

4. Results

4.1. Preprocessing and integration of tweets

Pre-processing resulted in 269,556 tweets from 119,611 Twitter users (out of which only 63 users had more than 100 tweets). This dataset represents 1331 counties from all 50 states, the District of Columbia, and Puerto Rico. The range of the ADI is from 3 to 98. Fig. 2 diagrams the pre-processing workflow. Table 1 summarizes the characteristics of the final dataset.

4.2. Topic modeling

We evaluated models ranging from 10 to 50 topics and selected the model with the highest coherence score, (coherence score 0.571) and 45 topics. Coherence scores for 10 to 50 topics are plotted in Supplementary Fig. 1. We named topics based on the common theme of the top words. For example, we defined topic 1 as "Shopping" due to top words "toilet", "paper", "store", "shop", "walmart", and "groceri" (stemmed version of groceries). The top 10 words in each topic are shown using word clouds in Fig. 3, wherein the font size in each plot reflects the importance of a word in a specific topic. Representative tweets (tweets with the highest probability of belonging to the given topic) for all topics are available in Supplementary Table 1.

4.3. Comparing topic prevalence over time

We present the topic-dynamics from January to March including the average distribution of topics that peaked by month. For each month, topic prevalence compared to both of the other months had a significance of $p < .0001$ unless indicated otherwise.

In January (Fig. 4), there were significant peaks in topics such as intense expression, negative expression, and personal expression (vs. Mar, $p < .001$). These topics are associated with profanity, anxiety, and emotions. There was also a peak in discussion regarding an early understanding of the novel disease, namely symptoms, flu deaths, and preventative measures (vs. Feb, $p < .01$; vs. Mar, $p < .05$). Further, there was significant discussion regarding China, international outbreak events (vs. Feb, $p < .01$), and ethnicity, as well as tweets concerning case counts (vs. Feb, $p < .05$), hotspots (vs. Feb, ns), and confirmed cases.

In February (Fig. 5), there was a significant rise in the discussion

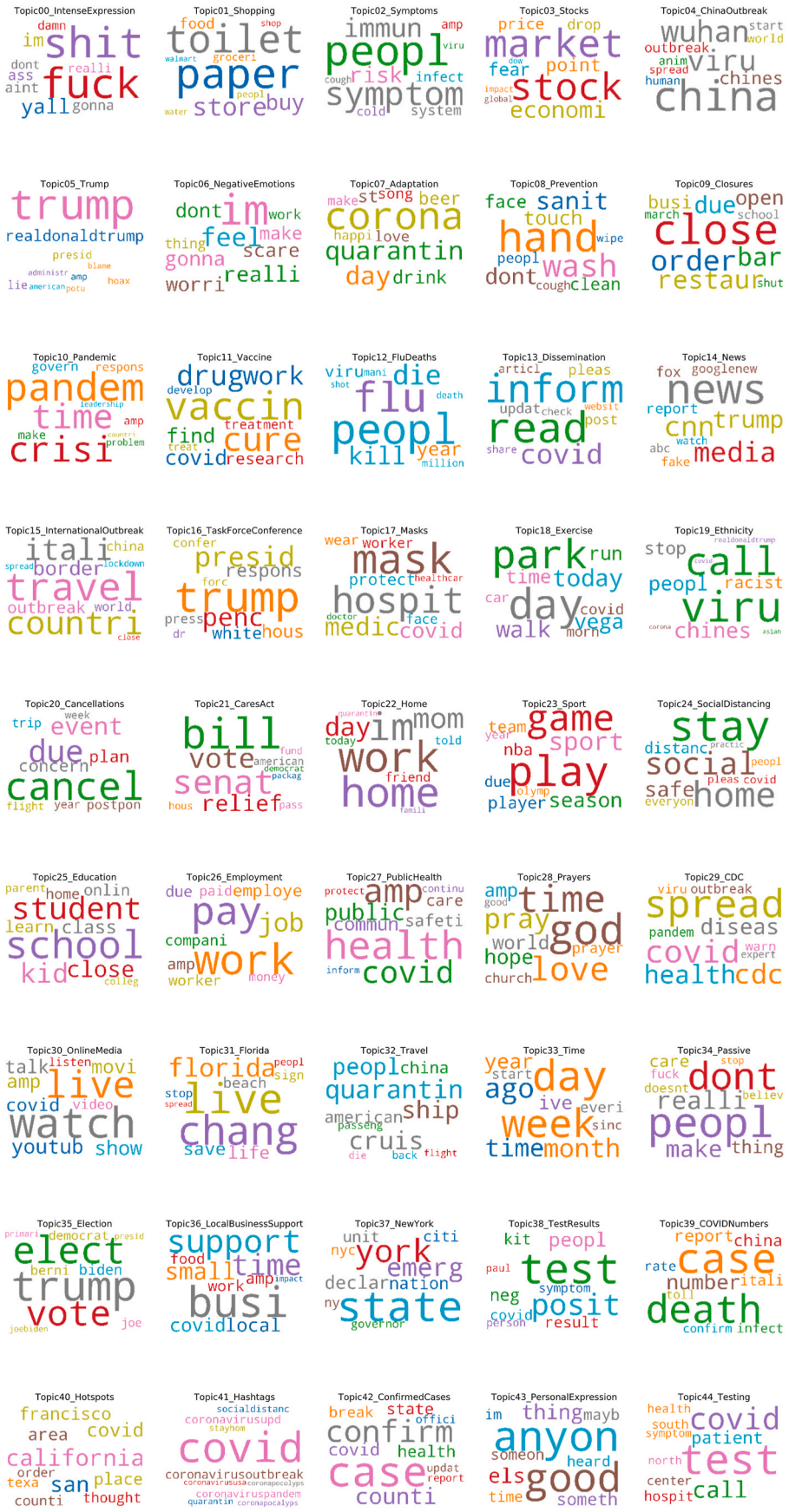

Fig. 3. Visualization of the top 10 words in all topics.

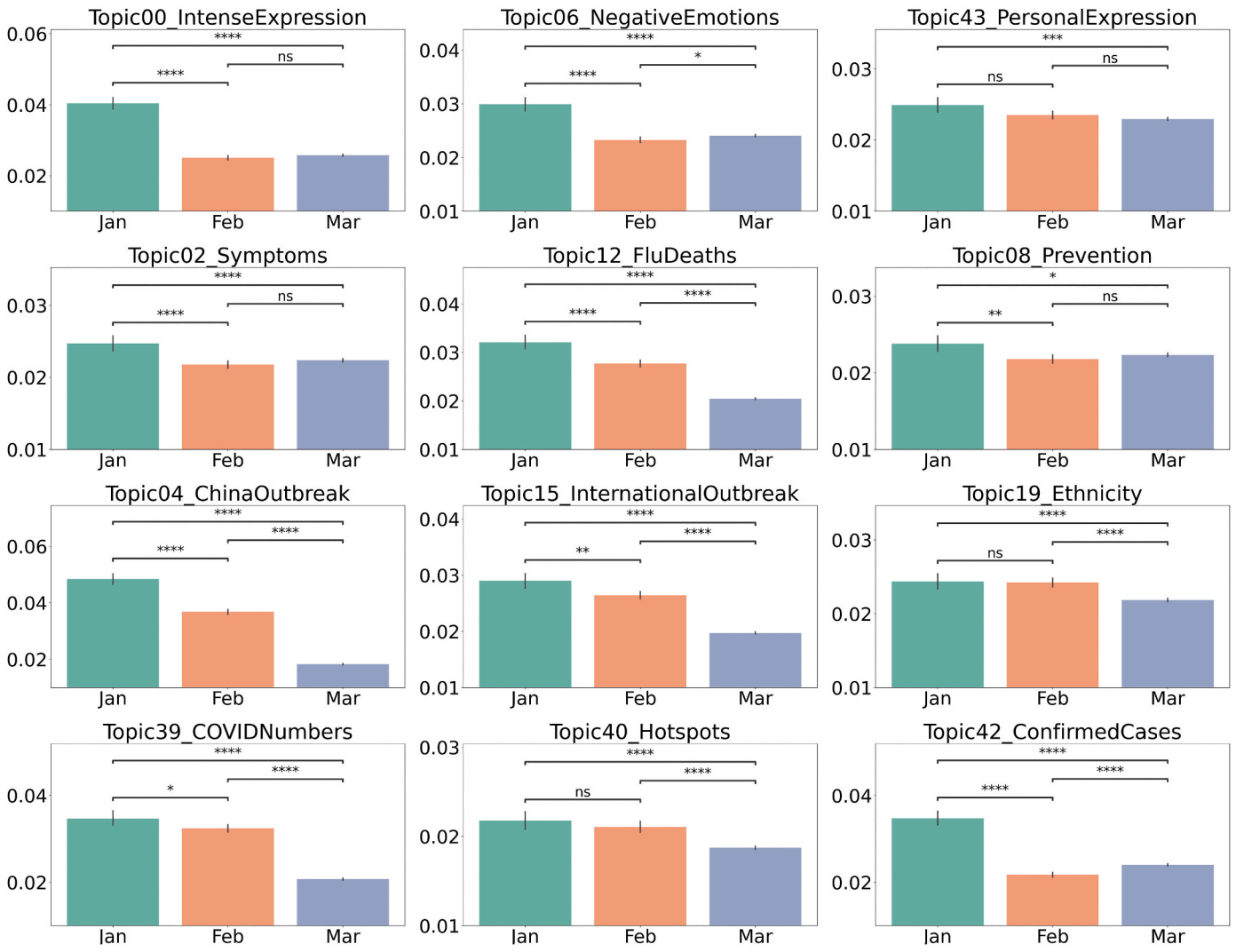

Fig. 4. Distribution of topics with higher proportions in tweets posted in January. Topics that had the same proportions for all months not shown. Significance testing results from two-sided Welch's *t*-test with Bonferroni correction. Significance legend: ns: 5.00e-02 < p ≤ 1.00e+00. *: 1.00e-02 < p ≤ 5.00e-02. **: 1.00e-03 < p ≤ 1.00e-02. ***: 1.00e-04 < p ≤ 1.00e-03. ****: p ≤ 1.00e-04.

surrounding the election, President Trump, news articles, stocks, the task force conference, and the CDC (Centers for Disease Control and Prevention). February also saw a significant discussion surrounding vaccines and travel (vs. Jan, $p < .05$).

In March (Fig. 6), there was a rise in discussions related to social distancing and disease mitigation strategies, namely closures, cancellations (vs. Feb, $p < .001$), social distancing, staying home, online media (vs. Jan, $p < .05$), and education. In general, there were higher topic proportions of activities related to quarantine, in particular exercising, sport, shopping, prayers, words related to time, and adaptation. March also resulted in more dissemination of information, discussion regarding the CARES Act, discussion of cases in Florida and New York, and tweets related to employment and local business support. Finally, in March there was a significantly higher proportion of tweets related to the pandemic (vs. Feb, $p < .001$), public health measures, tests and test results, and also a higher prevalence of COVID-related hashtags.

4.4. Comparing topic prevalence between low and high ADI areas

The ADI-specific analysis revealed significant differences in topic prevalence between low and high ADI areas. Comparing areas at the highest and lowest quintiles of ADI designation demonstrated differential effects ($p < .001$) in tweets by county-level socioeconomic resourcing. Topics that are more likely to dominate in high ADI (lower resourced) counties and low ADI (higher resourced) counties are shown in Fig. 7A. Topic prevalence comparisons between low and high ADI designated tweets had a significance of $p < .0001$ unless indicated otherwise. Tweets from high ADI areas are more likely to share emotional content with intense, negative ($p < .01$), personal expression ($p < .01$) or prayers ($p < .05$), as well as news regarding confirmed cases, the outbreak in China, flu deaths, and the CARES Act. On the other hand, tweets from low ADI areas were more likely to discuss the impact of COVID-19 on hotspots, local businesses, and New York status. Topics related to the larger public health crisis ($p < .001$) and pandemic ($p = .001$), as well as dissemination of information, stocks ($p < .01$), and the task force conference ($p = .01$), were also significantly more prevalent in tweets from lower ADI areas. These areas were also more concerned about the progress of potential treatments like vaccines ($p < .001$). While tweets with political topics about elections ($p = .937$) and President Trump ($p = .605$) were more likely to come from low ADI areas, the differences were not statistically significant.

Observing the topic proportion progress from January through March (Fig. 7B), we noticed that "Intense Expression" and "CARES Act" topics had consistent trends at both high and low ADI areas, with the high ADI areas having an overall higher daily average topic probability. Furthermore, topics associated with public health policies and disease

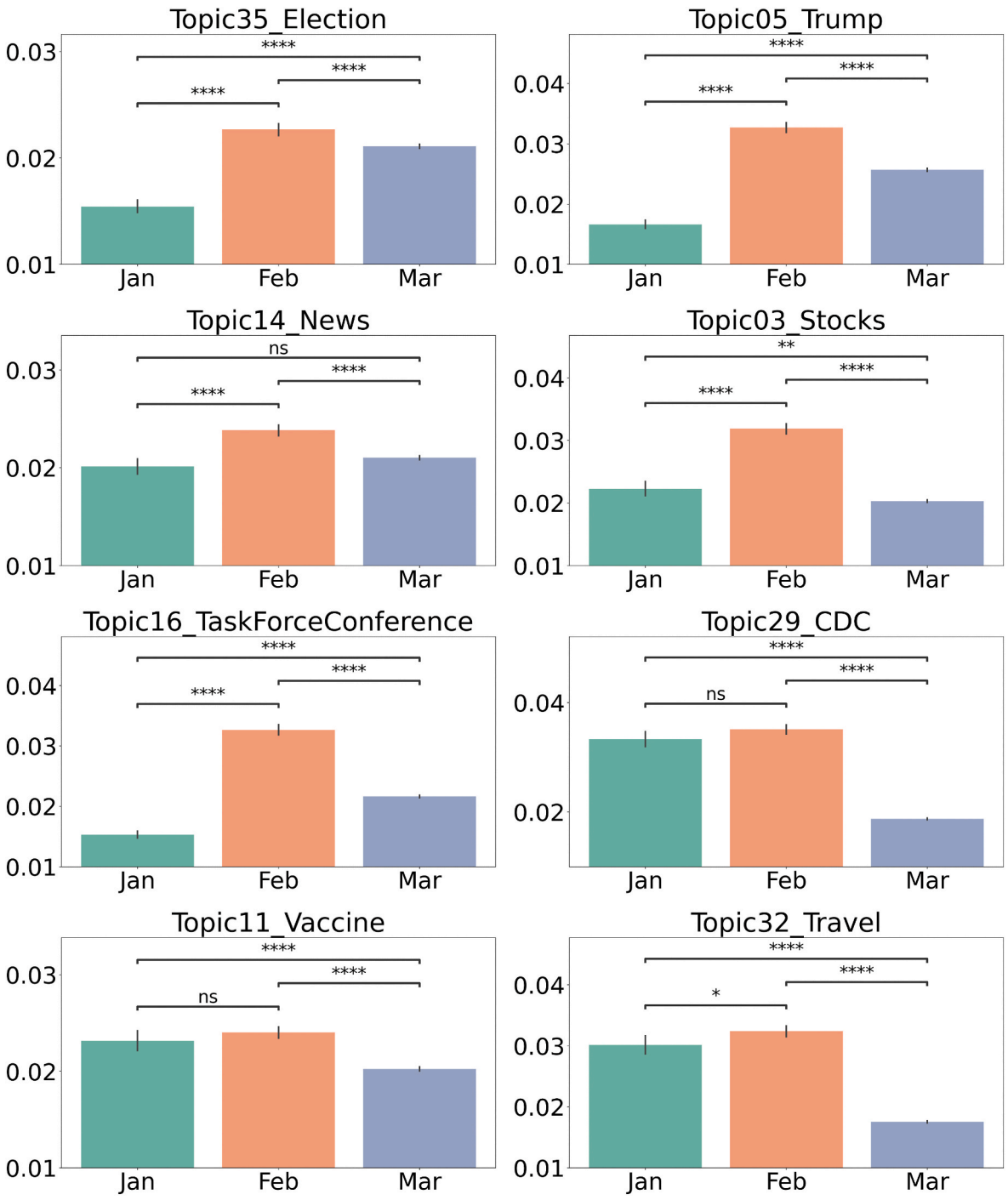

Fig. 5. Distribution of topics with higher proportions in tweets posted in February. Topics that had the same proportions for all months not shown. Significance testing results from two-sided Welch's *t*-test with Bonferroni correction. Significance legend: ns: 5.00e-02 < p ≤ 1.00e+00. *: 1.00e-02 < p ≤ 5.00e-02. **: 1.00e-03 < p ≤ 1.00e-02. ***: 1.00e-04 < p ≤ 1.00e-03. ****: p ≤ 1.00e-04.

mitigation strategies in March such as "Social Distancing" and "Local Business Support" arose in tweets from low ADI areas at a higher prevalence than tweets from high ADI areas.

4.5. Comparing topic prevalence between hotspots and non-hotspots

There were significant differences in the dominant topics between hotspots and non-hotspot areas. Tweets from hotspots were more likely to include topics relating to New York, social distancing, public health and pandemic, information dissemination, exercise/sport, education, time, closures, and employment (Fig. 8). Tweets that were not posted from hotspots were more likely to include topics pertaining to negative or intense emotion, concern regarding the CDC guidelines and task force conference, international events and flu deaths, as well as stocks and shopping.

4.6. Comparing topic prevalence within hotspots between low and high ADI areas

Comparing the topic prevalence of the within-hotspots-tweets between areas of high ADI and low ADI demonstrated that topics including confirmed cases, closures, intense expression, and hashtags were more

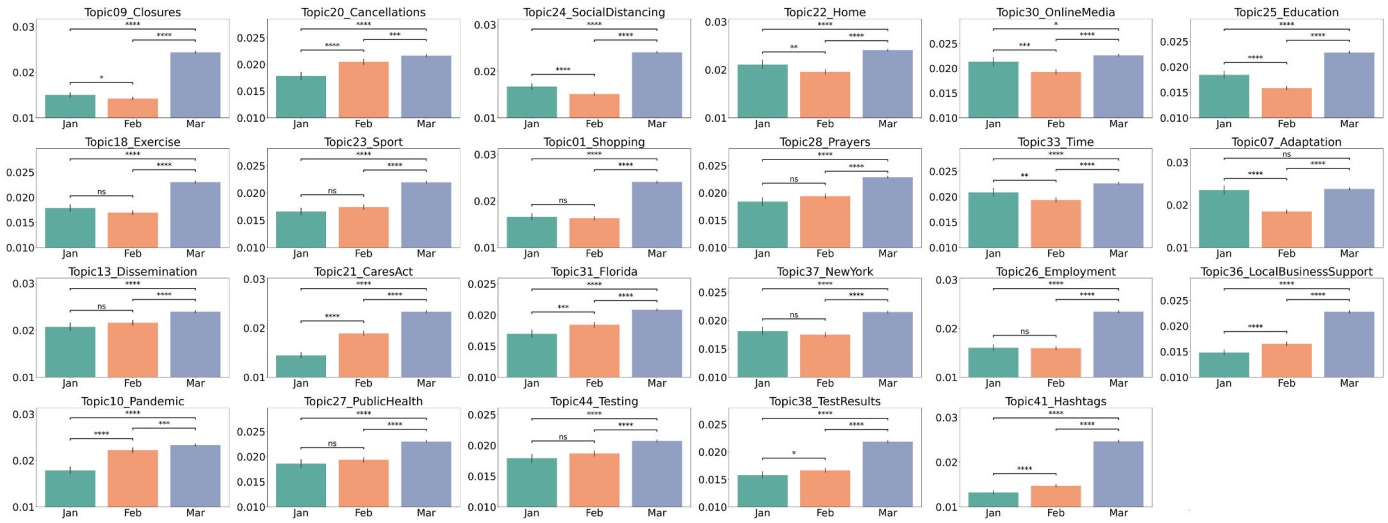

Fig. 6. Distribution of topics with higher proportions in March. Topics with same proportions for all months not shown. Significance testing results from two-sided Welch's *t*-test with Bonferroni correction. Significance legend: ns: 5.00e-02 < p ≤ 1.00e+00. *: 1.00e-02 < p ≤ 5.00e-02. **: 1.00e-03 < p ≤ 1.00e-02. ***: 1.00e-04 < p ≤ 1.00e-03. ****: p ≤ 1.00e-04.

prevalent from high ADI hotspots (Fig. 9A). Notably, tweets regarding employment concerns were also more likely to come from high ADI hotspots ($p < .001$), which wasn't significant in the previous analysis comparing ADI and hotspots separately. Tweets from low ADI hotspots were significantly more concerned with exercise, stocks, information dissemination, vaccine treatment, and cases in New York. We next observed the topic dynamics for selected topics from tweets collected in March (note that no high ADI areas were hotspots in January and February) (Fig. 9B). There were notable spikes in employment concerns and intense expression from high ADI hotspots, whereas these topics remain consistent throughout the month for tweets from low ADI hotspots. Tweets about New York and social distancing remained consistently high in low ADI tweets throughout March.

4.7. Chi-squared findings

Table 2 shows the Chi-squared testing results for our hypotheses. Testing for the differential concerns across less-resourced areas (low ADI) and high-resourced areas (high ADI), we found that the dominant topics differ significantly across areas with different socioeconomics levels ($p < .01$). Similarly, testing for differential concerns across hotspots and non-hotspots, we found that the dominant topics differ significantly relative to the pandemic severity ($p < .01$).

5. Discussion

Our analysis of COVID-19-related social media content demonstrates that Twitter can be used effectively to identify individual-level responses to infectious disease outbreaks in such a way that considers the impact of local-level socioeconomic resources and disease incidence. It shows too that socioeconomic disparity is associated with differential responses to the current COVID-19 pandemic, even among areas which are most severely impacted by disease cases. To our knowledge, this is the first study to link geocoded tweets to the ADI in order to explore the impact of geographic area-based socioeconomic status on tweet content.

This analysis follows the early pandemic timeline and establishes that topic modeling performs well in identifying major subjects of discussion on Twitter and successfully capturing the nuances of their variability. Though topic modeling has been applied to COVID-19-related tweets in an overlapping window of time (January 23 to March 7, 2020) [51], limited topics were identified and no analysis was reported about the emergence of new topics during that period. As the first cases of COVID-19 broke news in January, we found the fear sentiment in tweets as people were broadly focused on disseminating as much information as possible and similar conclusions were reported by Xue et al. [51]. As time progressed, there was increasing focus on local cases and events, public health information dissemination and testing, and quarantine activities.

Ordun et al. [52] explored topic prevalence over time in COVID-19 related tweets, however, the analysis was limited to reporting trends and lacked extended investigations of linking the trending topics to other health or social factors. In our study, by linking topic prevalence to socioeconomic status, we found that tweets from high ADI areas were more likely to share content regarding personal experiences, which ranged from positive affirmations of hope and prayers to negative or intense expressions of anxiety or frustration. This was not surprising given that the disparate impact of the pandemic and the associated economic fallout have, as predicted, disproportionately impacted poorer communities [53]. Furthermore, centuries of structural racism in the United States have led to lower resourcing in these areas and higher rates of medical co-morbidities that have been shown to increase COVID-19 risk [53] – all potentially contributing factors to an increase in intense, negative, and personal discussion in these areas pertaining to the public health and economic crisis.

Tweets from low ADI areas in March showed more discussion of social distancing and local business support, as quarantine policies hurt local businesses and resulted in discussions about bill relief to support these businesses. This result is consistent with the quicker response to stay at home orders from low ADI areas and is in line with recent reports of movement dynamic differences between low-income and high-income areas [54]. The higher prevalence of discussion surrounding stocks that was noted in low ADI areas was consistent with a greater stock market wealth residing amongst the wealthiest US households [55].

In the comparison between low and high ADI area hotspots, we identified that tweets with intense expression and those about employment insecurity were significantly more likely to come from high ADI hotspots. This reinforces the notion that, even after restricting to areas with high case counts, income and resource disparity result in disproportionate effects due to closures and job loss [56]. Furthermore, low ADI counties were significantly more concerned with information dissemination, cases in New York (on average a large low ADI hotspot), stocks, and vaccine treatment showing increased focus on social and institutional reactions to the crisis.

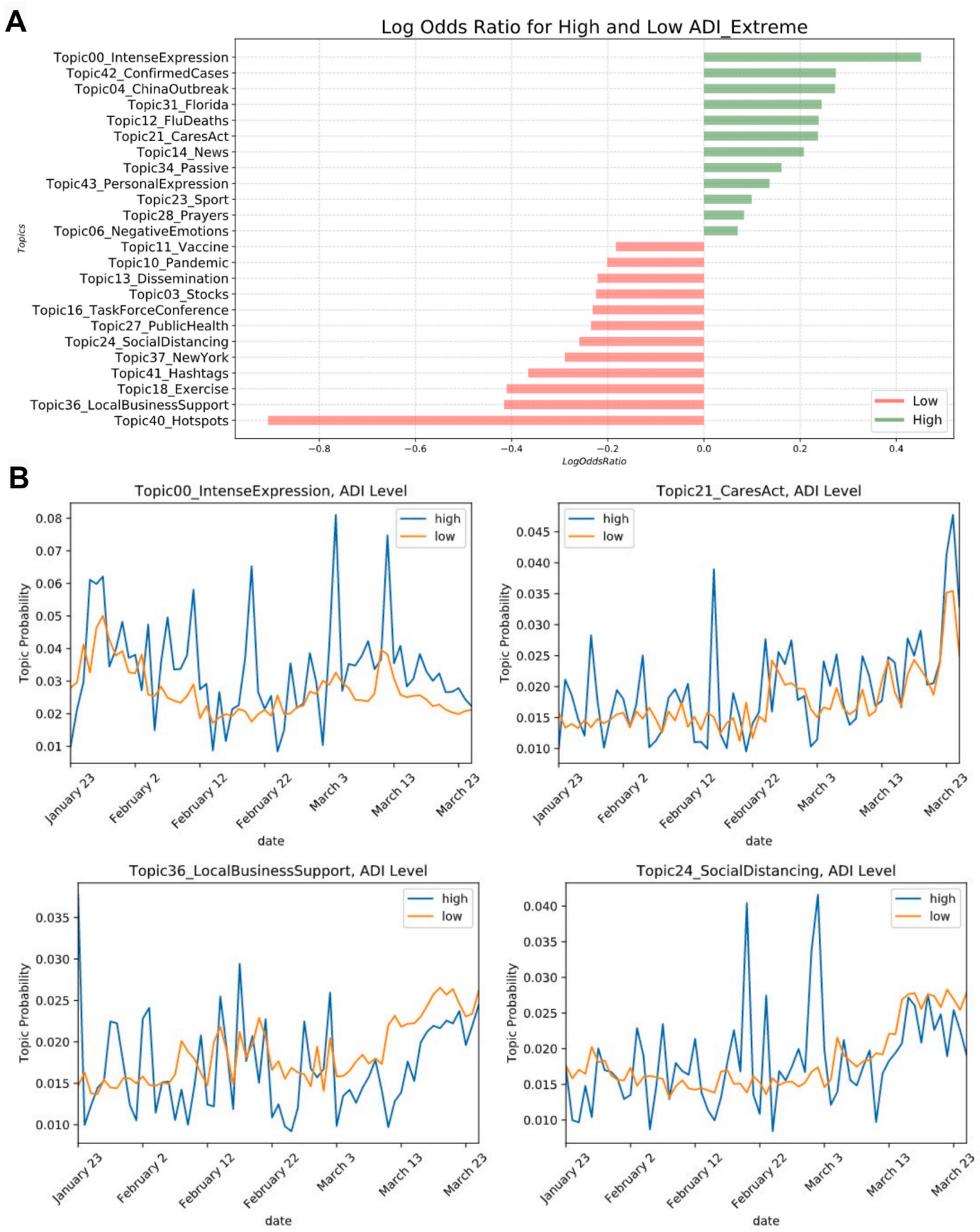

Fig. 7. Topic prevalence comparisons between High and Low ADI based on Log odds ratio. A. Topics with significant difference between both groups ($p < .05$) B. Topic dynamics for example topics.

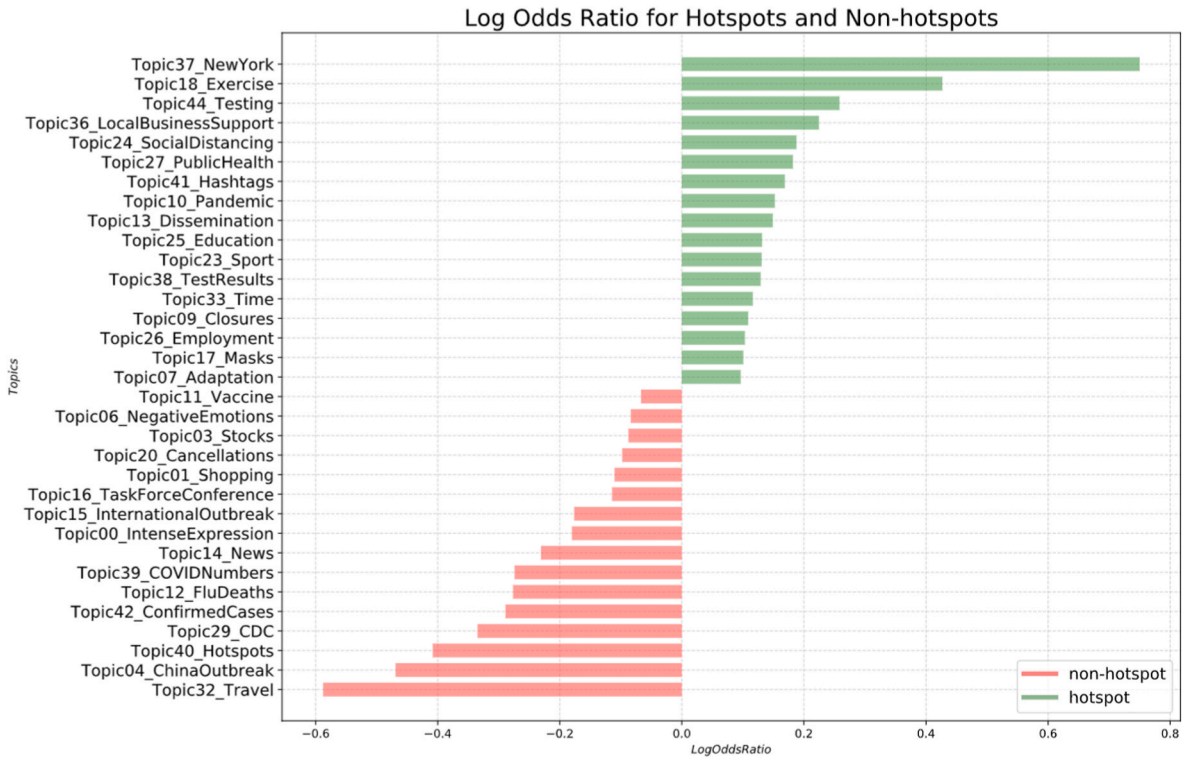

Fig. 8. Topic prevalence between hotpots vs non-hotspots based on log odds ratio.

Our approach of integrating a location-based socioeconomic index with Twitter topics offered increased insight into the topics inferred from the text, allowing a novel framework for assessing differential topics of conversation as they correlate to income, education, and housing disparities. Our integration of published COVID-19 hotspots further enables time-specific information of disease spread and how this corresponds to topics discussed on Twitter. These nuances are valuable for recognizing how public health communication, resource allocation policy, and information dissemination can respond to the needs of different communities, especially those with the lowest health resourcing, in future waves of the pandemic and emerging infectious disease outbreaks. Future public health efforts may use Twitter topic modeling to target messaging to the unique concerns of local communities and study the impact of health resource utilization. Our findings emphasize the importance of social media as a platform for public health communication as it is freely available to communities with different levels of socioeconomic resources. In fact, using public health communication to mitigate health disparities is not a novel concept [11], and is in line with future directions laid out in the National Institute on Minority Health and Health Disparities 2019 research framework [12]. However, the implementation of these methods should see expansion to other national institutions and organizations, such as the Office of Disease Prevention and Health Promotion and the Centers for Disease Control and Prevention. Furthermore, such initiatives need to be enhanced with more targeted messages, announcements and policies addressing the community level social and behavioral differences.

5.1. Limitations

Though our study successfully explored pandemic-related topics of conversation across tweets, there were a number of limitations, some of which have also reported in other studies [57]. One limitation is related to missing data. Due to data privacy, although Twitter data is publicly available, some tweets were posted from private accounts and thus could not be retrieved from the Twitter API. Another limitation that reduced the dataset sample size was that the Twitter Search API, which we used in this study, retrieves tweets from a reduced sample of all historic tweets posted about COVID-19. This sample is reduced further by focusing on English, US-specific, and geocoded tweets. Furthermore, due to restrictions with Twitter geocoding, we accepted some degree of positional inaccuracy in our study design, in that we were only able to collect geographic coordinates to the resolution of a county, and therefore characterized each tweet by the county rather than the census tract or block group. Given the inherent geographic masking techniques used by Twitter to promote confidentiality, and our study design which involved cross-area estimation and simple geographic centroid assessment [9], we acknowledge aggregation bias as a study limitation. However, previous work assessing the quality of deprivation indices shows that aggregated ADI is able to outcompete other metrics in capturing county and tract level information [58]. Furthermore, aggregated ADI has previously been used in other work to compare county-level socioeconomic status [59]. For our dataset, on average, the county ADI was distributed such that the median ADI was a reasonable approximation for the county. Finally, for technical reasons on our server, fewer tweets were scraped on some dates. However, we were still able to glean valuable conclusions from our data that represent the early pandemic progression.

6. Conclusion

Twitter analysis linking geocoded tweets to markers of geographic socioeconomic resourcing demonstrates that the COVID-19 pandemic has differentially impacted areas of the United States that are already institutionally underserved, even among areas most severely impacted. Highly-resourced areas were concerned with stocks, social distancing, and national-level policies, while low-resourced areas shared content with negative expression, prayers, and discussion of the CARES Act economic relief package. Within hotspots, increased discussion regarding employment in low versus high resourced areas was observed. This finding highlights the need to address the specific fears and concerns of these communities through personalized public health messaging at the local level. Our work indicates the emerging utility for

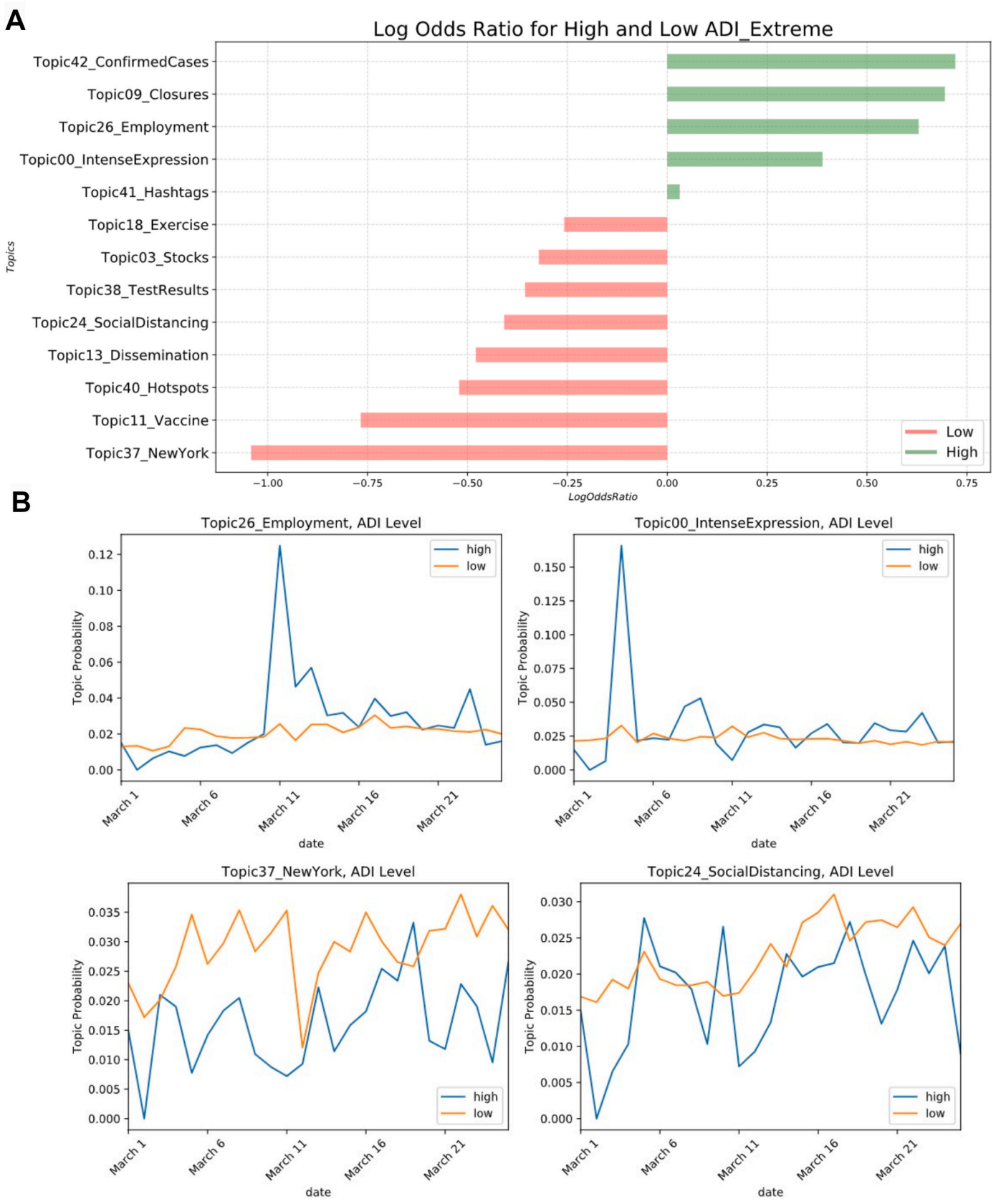

Fig. 9. Topic prevalence comparisons within Hotspots between low and high ADI areas. A. Topics with significant difference between the two groups ($p < .05$). B. Topic dynamics for example topics.

Table 2
Test results for (1) Dominant topics and ADI levels of each tweets, (2) Dominant topics and IsHotspots, and (3) ADI levels and IsHotspots.

Test Pairs	Chi-square Value	df	P-Value
Dominant Topics * ADI Levels	1660.841	88	<.001
Dominant Topics * IsHotspots	1399.751	44	<.001
ADI Levels * IsHotspots	18338.770	2	<.001

linking natural language processing techniques to real-time social media data and measures of social determinants of health. In future work, we plan to further analyze the sentiment of U.S. residents towards COVID-19 vaccination in areas with socioeconomic disparities. The speed at which vaccine-related misinformation is being propagated is alarming and has negative ramifications on global population health. We plan to investigate whether the volume and speed of misinformation differ relative to socioeconomic status and, specifically, if residents in less-resourced areas are disproportionately impacted by misinformation.

Funding

This research was supported in part by the Gruber Foundation (to A. V.).

Declaration of competing interestCOI

There is no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.104336.

References

[1] Centers for Disease Control and Prevention, If you are sick or caring for someone. https://www.cdc.gov/coronavirus/2019-ncov/if-you-are-sick/index.html, 2020. (Accessed 4 April 2020).
[2] Centers for Disease Control and Prevention, Social distancing, quarantine, and isolation. www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/social-distancing.html, 2020. (Accessed 4 April 2020).
[3] A. Wilder-Smith, D.O. Freedman, Isolation, quarantine, social distancing and community containment: pivotal role for old-style public health measures in the novel coronavirus (2019-nCoV) outbreak, J. Trav. Med. 27 (2) (March 2020).
[4] L. Buchanan, J. Patel, B. Rosenthal, et al., A month of coronavirus in New York city: see the hardest-hit areas. The New York times. https://www.nytimes.com/interactive/2020/04/01/nyregion/nyc-coronavirus-cases-map.html, 1 April 2020. (Accessed 20 July 2020).
[5] N. Chiwaya, J. Murphy, Tracking New Coronavirus Cases in the First Wave of Hot Spots across the United States, NBC News, 1 April 2020. https://www.nbcnews.com/health/health-news/coronavirus-count-state-day-2020-united-states-n1173421. (Accessed 20 July 2020).
[6] M. Chowkwanyun, A. Reed, Racial health disparities and Covid-19 – caution and context, N. Engl. J. Med. 383 (3) (2020) 201–203.
[7] R.A. Oppel Jr., R. Gebeloff, K.K. Lai, et al., The Fullest Look yet at the Racial Inequity of Coronavirus 25, The New York Times, 5 July 2020, p. 50. https://www.nytimes.com/interactive/2020/07/05/us/coronavirus-latinos-african-americans-cdc-data.html. (Accessed 17 July 2020).
[8] J. Younis, H. Freitag, J.S. Ruthberg, J.P. Romanes, C. Nielsen, N. Mehta, Social media as an early proxy for social distancing indicated by the COVID-19 reproduction number: observational study, JMIR Publ. Health Surveill. 6 (4) (2020), e21340.
[9] J. Kwon, C. Grady, J.T. Feliciano, S.J. Fodeh, Defining facets of social distancing during the COVID-19 pandemic: twitter Analysis, J. Biomed. Inf. (2020).
[10] X. Huang, Z. Li, Y. Jiang, X. Li, D. Porter, Twitter reveals human mobility dynamics during the COVID-19 pandemic, PloS One 15 (11) (2020), e0241957.
[11] V.S. Freimuth, S.C. Quinn, The contributions of health communication to eliminating health disparities, Am. J. Publ. Health 94 (12) (2004) 2053–2055.
[12] J. Alvidrez, D. Castille, M. Laude-Sharp, A. Rosario, D. Tabor, The national Institute on minority health and health disparities research framework, Am. J. Publ. Health 109 (S1) (2019 Jan) S16–S20.
[13] S. Afyouni, A.E. Fetit, T.N. Arvanitis, #DigitalHealth: exploring users' perspectives through social media analysis, Stud. Health Technol. Inf. 213 (2015) 243–246.
[14] A. Benetoli, T.F. Chen, P. Aslani, How patients' use of social media impacts their interactions with healthcare professionals, Patient Educ. Counsel. 101 (3) (2018) 439–444.
[15] F. Greaves, D. Ramirez-Cano, C. Millett, A. Darzi, L. Donaldson, Use of sentiment analysis for capturing patient experience from free-text comments posted online, J. Med. Internet Res. 15 (11) (2013) e239–e251.
[16] F. Alemi, M. Torii, L. Clementz, D.C. Aron, Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments, Qual. Manag. Health Care 21 (1) (2012) 9–19.
[17] Ranjitha Kashyap, Ani Nahapetian, Tweet Analysis for User Health Monitoring, 2015, pp. 348–351.
[18] Shirley Ann Williams, Melissa Terras, Claire Warwick, What people study when they study Twitter: classifying Twitter related academic papers, J. Doc. 69 (2013).
[19] W. Ahmed, P.A. Bath, L. Sbaffi, et al., Novel insights into views towards H1N1 during the 2009 Pandemic: a thematic analysis of Twitter data, Health Inf. Libr. J. 36 (1) (2019) 60–72.
[20] M. Odlum, S. Yoon, What can we learn about the Ebola outbreak from tweets? Am. J. Infect. Contr. 43 (6) (2015) 563–571.
[21] M.J. Paul, M. Dredze, D. Broniatowski, Twitter improves influenza forecasting, PLoS Curr. 6 (2014).
[22] S. Masri, J. Jia, C. Li, et al., Use of Twitter data to improve Zika virus surveillance in the United States during the 2016 epidemic, BMC Publ. Health (2019) 761.
[23] A. Sarker, S. Lakamana, W. Hogg-Bremer, A. Xie, M.A. Al-Garadi, Y.C. Yang, Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource, J. Am. Med. Inf. Assoc. 27 (8) (2020) 1310–1315.
[24] W. Ahmed, J. Vidal-Alaball, J. Downing, F. Lopez Segui, COVID-19 and the 5G conspiracy theory: social network analysis of twitter data, J. Med. Internet Res. 22 (5) (2020), e19458, 2020.
[25] G.R. Shinde, A.B. Kalamkar, P.N. Mahalle, N. Dey, Data Analytics for Pandemics: A COVID-19 Case Study, CRC Press, 2020.
[26] M.J. Horry, S. Chakraborty, M. Paul, A. Ulhaq, B. Pradhan, M. Saha, N. Shukla, COVID-19 detection through transfer learning using multimodal imaging data, IEEE Access 8 (2020) 149808–149824.
[27] H.R. Bhapkar, P.N. Mahalle, N. Dey, K.C. Santosh, Revisited COVID-19 mortality and recovery rates: are we missing recovery time period? J. Med. Syst. 44 (12) (2020) 1–5.
[28] K.H. Grantz, H.R. Meredith, D.A. Cummings, C.J.E. Metcalf, B.T. Grenfell, J. R. Giles, A. Wesolowski, The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology, Nat. Commun. 11 (1) (2020) 1–8.
[29] M.K. Chen, J.A. Chevalier, E.F. Long, Nursing home staff networks and COVID-19, Proc. Natl. Acad. Sci. Unit. States Am. 118 (1) (2021).
[30] N. Dey, R. Mishra, S.J. Fong, K.C. Santosh, S. Tan, R.G. Crespo, COVID-19: psychological and psychosocial impact, fear, and passion, Digit. Govern.: Res. Pract. 2 (1) (2020) 1–4.
[31] Search tweets API. https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets. https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets.
[32] Formal twitter vocabulary. https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/guides/basic-stream-parameters.
[33] J. Van den Broeck, S.A. Cunningham, R. Eeckels, K. Herbst, Data cleaning: detecting, diagnosing, and editing data abnormalities, PLoS Med. 2 (10) (2005) e267.
[34] Steven Bird, Edward Loper, Ewan Klein, Natural Language Processing with Python, O'Reilly Media Inc., 2009.
[35] A. Fan, F. Doshi-Velez, L. Miratrix, Assessing topic model relevance: evaluation and informative priors, Stat. Anal. Data Min.: ASA Data Sci. J. 12 (3) (2019) 210–222.
[36] Z.Y. Ming, K. Wang, T.S. Chua, Vocabulary filtering for term weighting in archived question search, in: M.J. Zaki, J.X. Yu, B. Ravindran, V. Pudi (Eds.), Advances in Knowledge Discovery and Data Mining. PAKDD 2010. Lecture Notes in Computer Science, vol. 6118, Springer, Berlin, Heidelberg, 2010.
[37] Geopy 2.0.0. https://pypi.org/project/geopy/. (Accessed 15 July 2020).
[38] University of Wisconsin School of Medicine Public Health, Area deprivation index v2.0. https://www.neighborhoodatlas.medicine.wisc.edu/, 2015. (Accessed 12 May 2020).
[39] A.J. Knighton, L. Savitz, T. Belnap, B. Stephenson, J. VanDerslice, Introduction of an area deprivation index measuring patient socioeconomic status in an integrated health system: implications for population health, EGEMS (Washington DC) 4 (3) (2016 Aug 11) 1238.
[40] P. Vart, J. Coresh, L. Kwak, S.H. Ballew, G. Heiss, K. Matsushita, Socioeconomic status and incidence of hospitalization with lower-extremity peripheral artery disease: atherosclerosis risk in communities study, J. Am. Heart Assoc.: Cardiovasc. Cerebrovasc. Dis. 6 (2017).
[41] Data from the New York Times, based on reports from state and local health agencies. https://github.com/nytimes/covid-19-data. (Accessed 1 May 2020).
[42] Xun Li, Qinyun Lin, Marynia Kolak, The U.S. COVID-19 Atlas. https://www.uscovidatlas.org, 2020. (Accessed 3 April 2020).
[43] D. Blei, A. Ng, M. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[44] E.S. Negara, D. Triadi, R. Andryani, Topic modelling twitter data with latent dirichlet allocation method, in: 2019 International Conference on Electrical Engineering and Computer Science (ICECOS), 2019, pp. 386–390.
[45] Andrew Kachites McCallum, MALLET: a machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.
[46] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., Scikit-learn: machine learning in Python, JMLR (2011) 2825–2830.
[47] J.M. Bland, D.G. Altman, Statistics notes: transforming data, Bmj 312 (7033) (1996) 770.

[48] S. Liu, Y. Wang, J. Zhang, C. Chen, Y. Xiang, Addressing the class imbalance problem in Twitter spam detection using ensemble learning, Comput. Secur. 69 (2017) 35–49.
[49] V. Prabhu, A.B. Rosenkrantz, Imbalance of opinions expressed on Twitter relating to CT radiation risk: an opportunity for increased radiologist representation, Am. J. Roentgenol. 204 (1) (2015) W48–W51.
[50] M. Delacre, D. Lakens, C. Leys, Why psychologists should by default use Welch's t-test instead of Student's t-test, Int. Rev. Soc. Psychol. 30 (1) (2017).
[51] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, T. Zhu, Public discourse and sentiment during the COVID 19 pandemic: using latent dirichlet allocation for topic modeling on twitter, PloS One 15 (9) (2020), e0239441.
[52] C. Ordun, S. Purushotham, E. Raff, Exploratory Analysis of Covid-19 Tweets Using Topic Modeling, Umap, and Digraphs, 2020 arXiv preprint arXiv:2005.03082.
[53] S. Galea, S.M. Abdalla, COVID-19 pandemic, unemployment, and civil unrest: underlying deep racial and socioeconomic divides, J. Am. Med. Assoc. 324 (3) (2020) 227–228.
[54] J. Valentino-DeVries, D. Lu, G.J.X. Dance, Location Data Says it All: Staying at Home during Coronavirus Is a Luxury, The New York Times, 3 April 2020. https://www.nytimes.com/interactive/2020/04/03/us/coronavirus-stay-home-rich-poor.html. (Accessed 3 April 2020).
[55] L. Ricketts, When the Stock Market Rises, Who Benefits? Federal Reserve Bank of St. Louis, 2018. https://www.stlouisfed.org/on-the-economy/2018/february/when-stock-market-rises-who-benefits. (Accessed 17 July 2020).
[56] N. Spievack, J. Gonzalez, S. Brown, Latinx unemployment is highest of all racial and ethnic groups for the first time on record, Urban Wire (2020). Accessed June 7, 2020.
[57] A. Joshi, N. Dey, K.C. Santosh (Eds.), Intelligent Systems and Methods to Combat Covid-19, Springer, 2020.
[58] B. Glassman, The multidimensional deprivation index using different neighborhood quality definitions, in: Prepared for the Western Economic Association Annual Conference, 2020.
[59] Mayo Clinic. County-Level Area Deprivation Index Scores and Quintiles by Year. Accessed March 6, 2021.