COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis

[1] P.Mohan, [2] Pradeep Burri, [3] Kanipakam Bhanu Moorty , [4] K.Vishnu Vardan Varma,
CSE Department,
[1,2,3,4] Assistant Professor, Dhruva Engineering Collage, Hyderabad.
Shree Engineering Collage, Hyderabad.

ABSTRACT

Misinformation about the COVID-19 virus is spreading rapidly online, and it might be very harmful. In this work, we apply machine learning to measure COVID-19 discourse among online anti-vaccination ("anti-vax") activists. The anti-vaccine movement, we have found, isCOVID-19 has received less attention from the anti-vaccination camp ("anti-vax") than vaccination itself.However, the anti-vax community displays a wider variety of "_favors" of COVID-19 topics, and so can appeal to a larger proportion of people looking for COVID-19 guidance online, including those who are leery of a mandatory fast-tracked COVID-19 vaccine and those who are interested in alternative treatments. Therefore, it seems that the anti-vaccine movement will be more successful in attracting new supporters in the future than the pro-vaccine movement will. This is worrying since it means that the globe will fall short of giving herd immunity against COVID-19, leaving nations vulnerable to future resurgences of the disease. We provide a mechanistic model that provides insight into these findings and may be useful for gauging the potential efficacy of intervention efforts.Our method is salable, so it may be used to the pressing issue of social media platforms needing to sift through vast amounts of online health misinformation and deception.

I. INTRODUCTION

Experts in the scientific community believe that creating a vaccination is crucial to successfully combating COVID-19. This, however, presupposes that a sufficient number of individuals would really become vaccinated.in order to increase the overall population's resistance. In order to ensure herd immunity, high vaccination rates against COVID-19 among the younger population are necessary [1] since vaccinations are less effective in the elderly. But even with current vaccines, such as those against measles, there is significant resistance, with some parents choosing not to vaccinate their children. A greater proportion of people in the United States and elsewhere contracted measles this year because of vaccination opposition [2]. Anticipate comparable resistance to any potential future COVID-19 vaccination [3, 4]. Vaccinating all school-aged children against COVID-19 might lead to a worldwide conflict in public health. Therefore, it is crucial for scientists, public health practitioners, and governments to have a deeper understanding of such objections before a COVID-19 vaccine is released.Vaccine sceptic have found a welcoming home in online social media, particularly the built-in communities that sites like Facebook (FB) provide (anti-vax)to meet together and spread (mis)information about health. This kind of false information is dangerous to people's health and safety [1, 4]. Similarly, pro-vaccine (pro-vax) advocates sometimes gather in similar online spaces to share information and lobby for official public health recommendations. Antivenin and pro-vax groups have been engaged in a heated online debate even before COVID-19. Misinformation regarding official medical advice and general scepticism about the state, the pharmaceutical sector, and emerging technologies like 5G communications are common themes in the stories told by anti-vaccine activists [1, 4, 5]. Adding fuel to the fire, the advent of the COVID-19 "infodemic" in January 2020 has led to a flood of false information about the virus on social media, some of which poses a serious risk to human life [6]. Harmful "cures" including drinking _sh tank additives, bleach, or cow urine have been suggested, and public health officials like the head of the United States' National Institute of Allergy and Infectious Diseases, Dr. Anthony Faucet, have been targeted in a concerted effort [7]. In addition, it has been widely circulated that people of colour are naturally resistant to COVID-19.

These factors may explain why members of particular minorities are more likely to be stereotyped as victims than others. The city of Chicago andIn early April 2020, African Americans accounted for _70% of deaths in Louisiana, although making up just _30% of the population [8, 9]. Moreover, the global community has seen a worrisome increase of COVID-19 weaponization against the Asian minority [10] [12]. It is also obvious that widespread acceptance of such falsehoods is not limited to a small subset of the population.In fact, despite assurances from infectious disease specialists, a recent Pew research [13] indicated that _30% of Americans think the COVID-19 virus was likely generated in a laboratory.

Regulating health-related falsehoods on social media platforms is difficult due to the flood of fresh material and the rapidity with which it spreads., [15]. Because of the isolation caused by the COVID-19 epidemic, individuals all around the globe are spending more time than usual on social media. Because of this, individuals are more likely to be exposed to hazardous COVID-19 treatments, cures, and lies, putting themselves and their connections at risk.

II. DATA AND MACHINE LEARNING ANALYSIS

Since each Facebook Page is a collection of individuals, the words "Facebook Page" and "cluster" may be used interchangeably to describe the concept being discussed. What are often referred to be "Fan Pages" or "Publicaccounts that represent groups, movements, communities, or public figures. A Facebook page's content is visible to anybody who visits the page, as stated in the terms of service. [Cite Section 21.5.5] It's important to note that a Facebook Page is distinct from a user's regular profile. Accounts belonging to people are assumed to be more discreet since they are intended just for their friends and family. No individual user data was analysed for this article. Our technique is based on [19] and [20], and it consists of examining the public content of Facebook Pages for anti vaccination ("anti-vax") and pro-vaccination ("pro-vax") groups. Following a manual identification of a seed set of sites on the topic of vaccinations, public policy around vaccinations, or the pro-vs.-contrary vaccination discussion, we use a snowball technique to collect the publicly accessible material of these online communities. Finally, their links to other fan sites are catalogue. Each subsequent phase involves an evaluation of the newly generated clusters using a hybrid of human coding and computer-assisted _alters.

We evaluated the postings and the 'about' part of each cluster to determine if it was (1) anti-vax or pro-vax and (2) included COVID-19 information. Vaccine supporters and opponentsIn order to be classified as pro-vax or anti-vax, either (a) at least 2 of the most recent 25 posts needed to address the topic, or (b) the page's title or "about" section needed to include that information. Each cluster was categorised by at least two researchers. If they couldn't agree on how to categorise the postings, a third researcher looked them over, and then the three of them talked it over. In every instance, an agreement was made. By doing so, we were also able to tell the difference between stuff that was meant to be taken seriously and that which was purely satirical. Facebook Pages have a natural propensity to screen out spam and phoney profiles. While we limited our analysis to the English language for this research, the same method may be utilised for any language. Additionally, we did not restrict our research to a certain geographical area.

Machine learning was used to examine the information that had been gathered into these clusters, first for the anti-vax group and subsequently for the pro-vax community. In particular, we analysed the development of subjects related to COVID-19 using Latent Dirichlet Allocation (LDA) [22], an unsupervised machine learning approach. The LDA approach models both document and topic distributions (in terms of words) to get a better understanding of the relationship between the two. Training the model involves tailoring these distributions to the data used in the model's creation. Wiki accurately describes LDA as [23] "[quote].. a generative statistical model that permits sets of observations to be explained by unobserved groups that explain why certain sections of the data are similar.

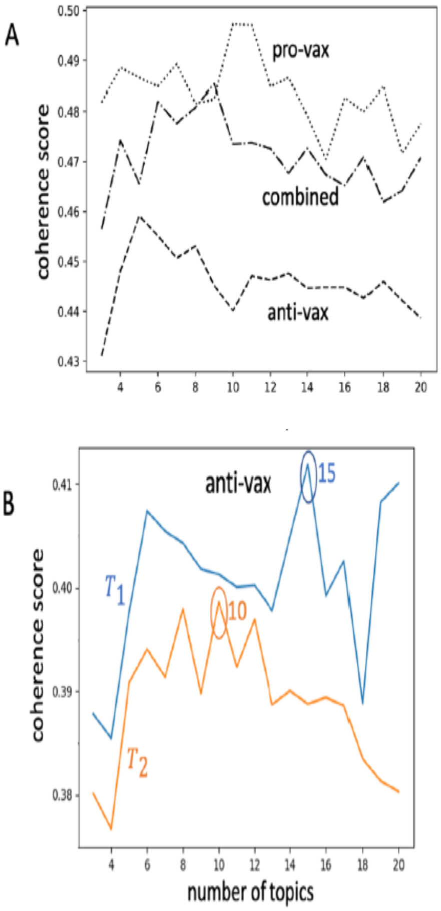
If, for instance, observations are words that have been compiled into papers, then this theory holds that these documents are comprised of a finite number of different themes and that the existence of any given word indicates that the document contains that term.is related to the subject matter of the paper. Machine learning and, more generally, artificial intelligence include LDA as one of its subject models. "The coherence score is a numerical indicator of how well the words in a text fit together in relation to a certain subject (see [22]). It's produced by applying a different algorithm to a pre-trained LDA model. The sum of a model's operaticcoherence is its overall coherence score. Per-topic coherence may be assessed using a variety of coherence measures. Based on a sliding window, one-set segmentation of the most frequently used terms, and an indirect con formation measure using normalized point-wise mutual information and the cosine similarity, we apply CV. It's a trove of probability measurements for how often topical key terms come together in instances. For a detailed analysis and explanation of CV, please see [22].

III. RESULTS

In this article, we concentrate on the internal evolution of COVID-19 during the early stages of the worldwide pandemic, before the first of recorded cases of COVID-19 in the United States.the date of one's demise is set for February 29, 2020 [25]. Therefore, we collected information from public Facebook posts from January 17, 2020, to February 28, 2020. This time frame was broken into sub-timeframes for the purpose of tracking evolution through time. Since we are only interested in the change over time and having more time intervals will result in lower quantities of data inside each and hence more _fluctuation, two intervals of equal length, T1 and T2, were selected. The first time period, from January 17, 2020, to February 7, 2020 (T1), includes 774 pro-vax posts and comments and 3,630 anti-vax posts and answers. There are a total of 673 pro-vax posts and responses and 3200 anti-vax posts and answers in the second time period (2/7/2020-2/28/2020, or T2). As a result, we have about the same quantity of information in each of our equally sized time periods. We validated that our findings hold up rather well over a variety of time frame selections.

Time period T1 generally corresponds to the time when COVID-19 was mostly perceived as a concern in Asia, and time period T2 roughly corresponds to the time when it became a severe issue in Europe. We further validated that the data split is identical for mentions of COVID-19 in article counts from global anglophone newspapers and worldwide Google trends to ensure that our data was reflective of the COVID-19 discourse throughout these periods.

Training sets for the LDA models included anti-vaccination posts from Time Period 1 and Time Period 2, pro-vaccination posts from Time Period 1, and pro-vaccination posts from Time Period 2.updates to T2 threads. Ten LDA models were trained for each datasets, with the number of topics parameter varying from 3 to 20, for a total of 180 models across the four categories.The Appendix has further information. We next ran the CV coherence method on these models, averaging the coherence scores across all topic counts.Plots of these mean scores are shown in Figures 1B and 1C.Figure 1A displays the outcome of applying the same method to both the anti-vaccination posts and the pro-vaccination posts in our datasets.



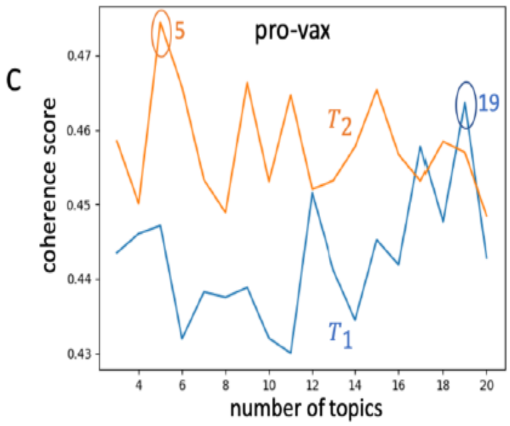


Figure 1 shows the estimated coherence scores CV for (A) anti-vaccine material (dashed line), pro-vaccine content (dotted line), and anti-vaccine mixed with pro-vaccine content (dashed-dotted line), across the whole research period (T1 CT2). (2) Anti-vaccineinformation for both the first (blue) and second (red) time periods (orange line).

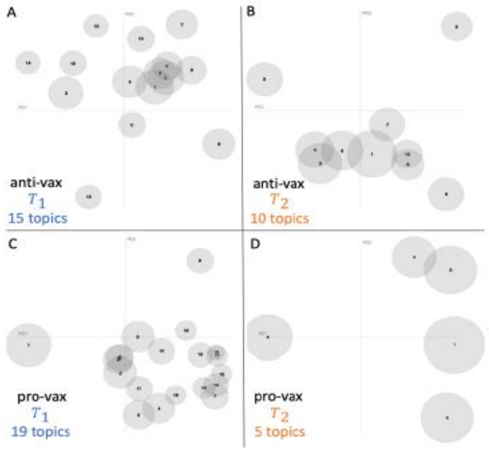
The ideal number of subjects, as determined by the highest possible coherence value (CV), is given. From T1 to T2, the ideal number of anti-vax issues decreases from 15 to 10. (C) Vaccination-supporting materials published during T1 (blue line) and T2 (red line) (orange line). From time point T1 to time point T2, the optimum number of subjects for pro-vaccine arguments decreases from 19 to 5.In Fig. 1A, the coherence score CV is bigger for the pro-vax side across all issues than it is for the anti-vax side, indicating that the former contains a more concentrated discussion of COVID-19 than the latter. This is in line with the prophylaxes community's uniform discussion of public health, in which the emphasis is placed on encouraging individuals to adhere to the recommendations of medical experts.

The downside of this increased general coherence for the pro-vaccine group is that it leaves them less prepared to interact with the vast array of more nebulous and, frequently, more radical views.

stories about COVID-19 that have just been spreading online. This might be a serious setback for the pro-vax movement since it means they are less likely to catch the eye of the many various kinds of users who are increasingly venturing into this virtual realm in quest of a certain nuanced "_favor" of the COVID-19 story that strikes their fancy. As a result, these consumers could be swayed to support the anti-vaccine movement.

By comparing the curves of the coherence score across the number of topics for periods T1 and T2, Figures 1B and 1C show the development over time. For the pro-vaccine group, the curve shifts from T1 to T2 (Fig. 1C), and the ideal number of topics drops significantly from 19 to 5. That fits with the theory that the provably group, unlike the anti-vax community, is trying to settle on a single interpretation and narrative of COVID-19.

While this may seem positive at first glance, it really indicates that support for vaccination and the pro-vaccine movement as a whole is waning.time to the many new users of varying backgrounds that are looking for their own personal COVID-19 story '_favor. However, the curves for the anti-vax group reveal a much smaller decrease in the ideal number of topics (from 15 to 10) and a downward trend from T1 to T2 (Fig. 1B), in contrast to the pro-vax community. Therefore, the anti-vax community is becoming more accepting of the increasingly heterogeneous population of new additions coming to the online health space by compensating a small increase in focus (reduction in the optimal number of topics) with an overall reduction in coherence, i.e., these 10 topics for T2 are effectively more blurry than the original 15 for T1.A more in-depth look at the relationships between the subjects and their informational distance from one another is shown in Figure 2 of this figure. The Davis package [26] is used to generate the resulting plot, which allows for a high-level overview of the subjects and their differences as well as a detailed examination of the words most strongly related with each topic. Insinuating a term's relation to a subject in this way is a new approach. However, as shown by the research in [26], simply ranking concepts according to their likelihood under a subject is not ideal for topic interpretation. You can learn all there is to know about Davis by reading [26].



This figure provides a visual representation of the relationships between the various subjects and the informational structure they share. The software Davis is used to generate this graphic. The circles in each graph represent the same themes as Fig. 1.maximum average coherence score; i.e., best possible range of subjects covered.

The two axes are main components of the distribution analysis, and their magnitude represents the marginal topic distribution, which is explored in depth in [26]. (Fig. 2D).Moreover, in Fig. 2B, the themes seem to be dispersed more evenly over the area than in Fig. 2D. These findings corroborate our earlier interpretations that the pro-vax community is more focused (or narrower) than the anti-vax community with regards to COVID-19 narratives, and that the pro-vax community is moving toward a common COVID-19 interpretation and narrative with less diversity on offer than the anti-vax community.

IV. TOWARD A MECHANISTIC MODEL INTERPRETATION

To further substantiate these empirical _endings and to provide a microscopic explanation for the results of machine learning, we developed a mechanical model. We created aA computer model of the online health discussion takes into account the interconnections of many different parts, or "components," each of which is defined by a vector x D (x1, x2,...), where each component xi signifies the strength of a certain aspect (such as government control) in the argument. It is not necessary to specify the precise nature of these components, such as whether they are words or short sentences. What's important is that there is a rich ecosystem of different building components. Despite its apparent simplicity, the mechanical model setup investigated in depth by Kate [1] reflects both the actual facts and the literature around the topics of online conversations of vaccine resistance. To simulate how this would work, we use a computer programme to pick and choose from among these elements. if the x-values of the components are sufficiently similar (i.e., homophobia in panel A), or dissimilar (i.e., heterotroph in panel B), then they will cluster together (or their clusters will cluster together, if they are already in a cluster) (i.e., heterotroph in panel B).

homophobia in Fig. 3A), or if they are sufficiently dissimilar (i.e. heterotroph in Fig. 3B). One dimension of our model's results is shown in Fig. 3. Weverified that a two-dimensional version yields the same findings, but with additional visual complexity due to the addition of the third dimension of time. Notably, it generates graphs that are aesthetically consistent with Fig. 2.

As illustrated in Figs. 1 and 2, the pro-vax community converges more quickly than in the case of homophobia (which is analogous to developing a more monolithic subject debate with few _favors). In contrast, as shown in Figs. 1 and 2, the heterotroph situation (which is analogous to constructing various subject conversations with multiple _savors, like the anti-vax group) takes more time to gel. The red dashed horizontal line in Figures 3A and 3B indicates the simulation stage that is generally consistent with Figures 2D and 2B for the pro-vax and anti-vax groups.

V. CONCERNS ABOUT THE STUDY

Several caveats exist with this research. Although Facebook is the most popular social networking site, there are several others worth investigating. Communities on every platform may be expected to exhibit similar patterns of interaction. It would be fascinating to compare our findings to, say, research centred on Twitter, where users tend to communicate via short, discrete utterances [17].Another issue is the potential impact of extraneous factors [16]. However, these online groups often have their own mechanisms in place to deal with bots and trolls. The content's particulars need to be examined in more depth. Since memes and photos are also circulated, we'll need to expand our search beyond text and maybe even LDA to do this. In addition, the time evolution of topics has to be compared explicitly with the output of the generative model. Across all platforms, the conclusions need to be refined into clear, actionable repercussions for policymakers, which requires further study. It is planned to get around these restrictions in further studies.

VI. CONCLUSION

These findings indicate that the anti-vaccine online community is more likely to be building a diversified and, as a result, widely accommodating conversation surrounding COVID-19 than the pro-vaccine online community.Therefore, the pro-vaccine community faces the danger of alienating the diverse ecosystem of new users who may participate in the online COVID-19 conversation, and who may bring a wide range of concerns, questions, and even disinformation and even lies with them.

This paper's study also represents a first step toward someday substituting for, or at least augmenting, the non-salable work of human moderators charged with spotting internet falsehoods. The mechanistic model (Fig. 3) can also be used to test out hypothetical situations to see how quickly coherence forms and what happens if the coherence surrounding certain topics is shattered, such as when people are discouraged from taking bleach or the even more recent 'COVID Organics' that are circulating as a cure in Madagascar, Africa, and elsewhere. To do this, we may use the empirical study in Fig. 2 performed over many consecutive time periods to track the rise in interest in potential new home remedy terms, such as "bleach." Then, instead of using generic language, social media platforms like Facebook might publish advertising that directly target these emerging concepts and terms.advancing official explanations from the medical establishment.

In sum, the results of this method demonstrate that the LDA algorithm, a machine learning technique, successfully identifies probable subjects within collections of postings from online communities discussing vaccines and COVID-19. Rather of relying on possibly biased, sluggish, and expensive human labeling, this method can handle massive amounts of data and generates results fast via the use of statistical grouping algorithms.

REFERENCES

[1] A. Kate, ``A postmodern Pandora's box: Anti-vaccination misinformationon the Internet,'' Vaccine, vol. 28, no. 7, pp. 1709_1716, Feb. 2010, dpi: 10.1016/j.vaccine.2009.12.022.

[2] L. Givetash, Global measles cases surge amid stagnating vaccinations.New York, NY, USA: NBC News, 2019. Accessed: Apr. 13, 2020.[Online]. Available: https://www.nbcnews.com/news/world/globalmeasles-cases-surge-amid-decline-vaccinations-n1096921

[3] B. Martin, Texas Anti-VAXes Fear Mandatory COVID-19 VaccinesMore Than the Virus Itself. Austin, TX,USA: Texas Monthly,2020. [Online]. Available: https://www.texasmonthly.com/news/texasanti- VAXes-fear-mandatory-coronavirus-vaccines/

[4] H. J. Larson, ``Blocking information on COVID-19 can fuel the spreadof misinformation,'' Nature, vol. 580, no. 7803, p. 306, Apr. 2020, dpi: 10.1038/d41586-020-00920-w.

[5] R. Schroeder and E. Laurie, Corona virus: Scientists Brand 5G Claims 'Com-plete Rubbish. London, U.K.: BBC News, 2020. Accessed: Apr. 5, 2020.[Online]. Available: https://WWW.bbc.com/news/52168096

[6] Corona virus Disease (COVID-19) Advice for the Public: Myth Busters,W. H. Organization, Geneva, Switzerland, 2020. Accessed: Apr. 13, 2020.[Online]. Available:https://www.who.int/emergencies/diseases/novelcoronavirus-2019/advice-for-public/myth-busters

[7] K. Bender and M. Shear, After Threats, Anthony Faucet toReceive Enhanced Personal Security. New York, NY, USA: TheNew York Times, 2020. Accessed: Apr. 2, 2020. [Online]. Available:https://WWW.anytime.com/2020/04/01/us/politics/corona virus-peculiarity.html

[8] S. Almasy, H. Yan, and M. Holcomb, Corona virus Pandemic HittingSome African-American Communities Extremely Hard. New York, NY,USA: CNN Health,

2020. Accessed: Apr. 13, 2020. [Online]. Available:HTTP://WWW.cnn.com/2020/04/06/health/us-corona virus-demimondaine/index.HTML

[9] A. Maqbool, Corona virus: Why Has The Virus Hit African Americans soHard. London, U.K.: BBC News, 2020. Accessed: Apr. 13, 2020. [Online]. Available: https://www.bbc.com/news/world-us-canada-52245690

[10] J. Guy. (2020). East Asian Student Assaulted in 'Racist' Corona virus Attack in London. Accessed: Apr. 13, 2020. [Online]. Available:https://www.cnn.com/2020/03/03/uk/coronavirus-assault-student-londonscli-intl-gbr/index.HTML

[11] H. Yan, N. Chen, and D. Naresh. (2020). What's Spreading FasterThan Corona virus in the US? Racist Assaults and Ignorant AttacksAgainst Asians. Accessed: Apr. 13, 2020. [Online]. Available:HTTP://WWW.cnn.com/2020/02/20/us/corona virus-racist-attacks-instantiate-Americans/index.HTML.

[12] M. Rajasthan, Korean Interpreter Says Men Yelling 'Chinese'Tried to Punch Her Off Her Bike. New York, NY, USA: Buzz Feed News, 2020. Accessed: Apr. 13, 2020. [Online]. Available: https://www.buzzfeednews.com/article/meghara/coronavirus-racismeurope-covid-19

[13] K. Schrieffer, Nearly Three-in-ten Americans Believe COVID-19 wasMade in a Lab, Washington, DC, USA: Pew Fact Tank, 2020. Accessed:Apr. 10, 2020. [Online]. Available: https://www.pewresearch.org/facttank/2020/04/08/nearly-three-in-ten-americans-believe-covid-19-wasmade-in-a-lab/

[14] R. Glengarry, The Corona virus is Stretching Facebook to its Limits.New York, NY, USA: CNN Business, 2020. Accessed: Apr. 13, 2020.[Online]. Available: HTTP://WWW.cnn.com/2020/03/18/tech/zuckerbergfacebook-coronavirus-response/index.html

[15] S. Frankel, D. Alba, and R. Hong. (2020). Surge of Virus Misinformation Stumps Facebook and Twitter. Accessed: Apr. 13, 2020. [Online].Available: https://www.nytimes.com/2020/03/08/technology/coronavirusmisinformation-social-media.HTML

[16] D. A. Brontosaur, A. M. Jami son, S. Q, L. Balalaika, T. Chen,A. Benton, S. C. Quinn, and M. Dredge, ``Weaponized health communication:Twitter bots and Russian trolls amplify the vaccine debate,''Amer. J. Public Health, vol. 108, no. 10, pp. 1378_1384, Oct. 2018, dpi: 10.2105/AJPH.2018.304567.

[17] Y. Lama, T. Chen, M. Dredge, A. Jami son, S. C. Quinn, andD. A. Brontosaur, ``Discordance between human papilloma virus Twitterimages and disparities in human papilloma virus risk and disease in theunited states: Mixed-methods analysis,'' J. Med. Internet Res., vol. 20,no. 9, Sep. 2018, Art. no. e10244, dpi: 10.2196/10244.

[18] T. Maria and S. Schoenberg, ``Thanks for your interest in our Facebook group, but it's only for dads: Social roles of Stay-at-Home dads,'' in Prof.19th ACM Conf. Compute.-Supported Cooperate. Work Social Compute.,2016, pp. 1361_1373, dpi: 10.1145/2818048.2819927.

[19] N. F. Johnson, R. Leafy, N. J. Freepost, N. Velasquez, M. Sheng,P. Manque, P. Dakota, and S. Touchy, ``Hidden resilience and adaptivedynamics of the global online hate ecology,'' Nature, vol. 573, no. 7773,pp. 261_265, Sep. 2019, dpi: 10.1038/s41586-019-1494-7.

[20] N. F. Johnson, M. Sheng, Y. Voronezh, A. Gabriel, H. Q, N. Velasquez,P. Manque, D. Johnson, E. Freepost, C. Song, and S. Touchy,``New online ecology of adversarial aggregates: ISIS and beyond,'' Science, vol. 352, no. 6292, pp. 1459_1463, Jun. 2016, dpi: 10.1126/science.aaf0675.

[21] Facebook Policies. (2020).Pages, Groups and Events Policies. Accessed:Apr. 13, 2020. [Online]. Available: https://www.facebook.com/policies/pages_groups_events