Topic Modeling of Twitter Data to Identify Issues Impacting Women and Girls' Education During the COVID-19 Pandemic

Arnab Dey, Nabamallika Dehingia, Anita Raj
Center on Gender Equity and Health, University of California San Diego, USA

Bharanidharan Radha Saseendrakumar
UCSD Health Department of Biomedical Informatics, University of California San Diego, La Jolla, CA

Tarang Tripathi
Department of Education Studies, University of California San Diego, USA

Background

The impact of the COVID-19 pandemic on education has been enormous. UNICEF reports that 188 nations have imposed countrywide school closures, affecting more than 1.6 billion students globally.[1] Evidence also indicates that these educational impacts may disproportionately affect girls, due to unequal domestic labor expectations and digital access gaps for online learning,[2] as well as increased risk for girl child marriage, adolescent pregnancy, and sexual exploitation — all of which have been exacerbated under the pandemic.[3,4] These impacts on girls' education are gender regressive and can create negative inter-generational effects across socio-economic and health dimensions.

A better understanding of these issues would be pivotal for tackling these enormous challenges and determining ways to prevent the extensive negative gendered consequences of the pandemic. Given the difficulties in collecting survey data during the pandemic, publicly available big data sources such as Twitter can be harnessed to identify issues that are being highlighted by organizations and individuals around the world. In this brief, we study Twitter conversations and use topic-modeling to identify priority issues around gender equity being discussed by global education organizations.

Our Approach

Sampling and data extraction: We identified three online platforms that curate international education organizations: 1) the Association of International Education Administrators, which focuses on international education leadership in higher education; 2) the National Association of Independent Schools, which supports kindergarten to twelfth grade schools on research, leadership, and development and; 3) Human Rights Careers, which highlights organizations dedicated to human rights including those related to human rights education and right to education organizations, for students from early childhood to young adulthood. We then screened all educational organizations identified via these platforms using the following inclusion criteria: a) having a Twitter handle; b) having more than 1,000 followers on Twitter; c) indicating commitment to international education in their Twitter description.

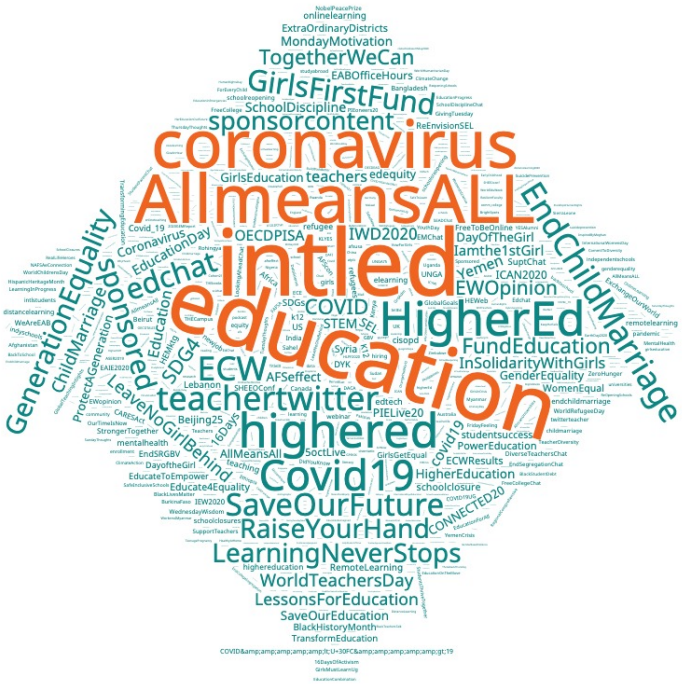
We identified 61 Twitter accounts of organizations working on education across national settings that met our criteria. Using the official Twitter API, we collected 61,995 tweets from these accounts made between March 01, 2020 and February 28, 2021 as our sample for these analyses.

Analysis: We extracted all hashtags in the sample and created a Word Cloud to understand the gender-focus in these tweets from the aforementioned educational organizations. To focus on issues related to gender equity in education, we further narrowed our sample to those with keywords related to women/girls or gender: "girl/girls," "gender," "woman/women," "marriage." Tweets without at least one keyword were excluded from further analysis. This process narrowed our sample to 9,058 tweets specific to gender issues in education. We then cleaned the tweets by removing non-informative text and characters such as URLs, tags of other Twitter accounts, special characters, punctuations, stop words, and emojis.

We implemented topic modeling with Latent Dirichlet Allocation (LDA)[5,6] using Python's Gensim wrapper for the MALLET toolkit for statistical natural language processing.[7,8] We selected ten topics to be used by the model as that number maximized the coherence score of the topics.[9,10] We then followed standard procedure for LDA[5,9] to assign one of those ten topics as dominant in each tweet in our sample. Two co-authors then examined the words corresponding to each dominant topic and assigned a qualitative annotation to the topics identified.[11] This methodology could be implemented on a country specific basis using organizations within that country that focus on education.

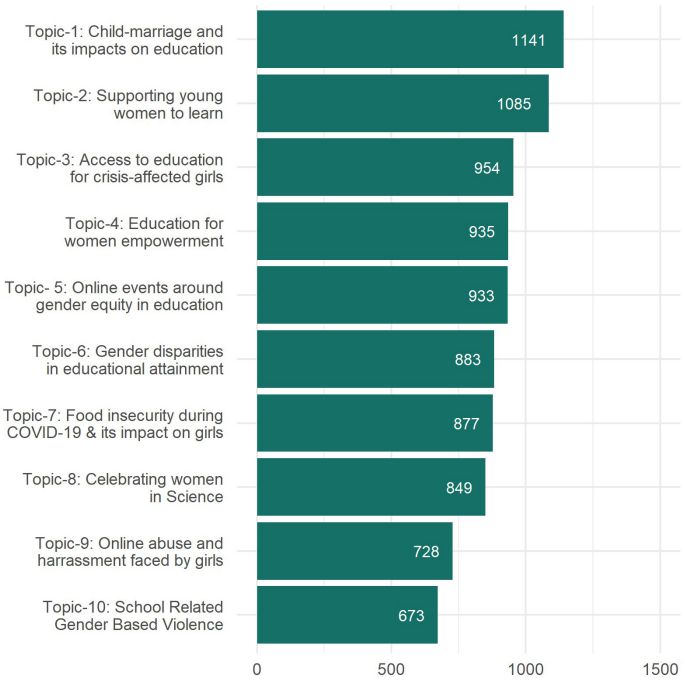
Results: Our initial analysis of hashtags across all identified tweets (N=61,995 tweets) found that the most tweeted hashtag was #COVID19, followed by #intled and #education (see Figure 1). Nonetheless, gender and equity focused hashtags were also seen; these included #AllmeansALL, #EndChildMarriage, #GirlsFirstFund, #LeaveNoGirlBehind, #InSolidarityWithGirls, suggesting a concern or emphasis related to the inter-relationship between girl's education and early marriage. Narrowing further, gender equity in education was the focus of 15 percent of all tweets (n = 9,058 tweets).

Figure 1: Word cloud of hashtags used in the sample tweets



With our LDA analysis we classified the sample of 9,058 gender equity and education tweets into ten different topics (Figure 2). These topics covered a wide range of issues related to education; the majority indicated concerns about girls' retention in schools under COVID-19 and risk for girl child marriage in the absence of school retention (Topics 1–3). Secondarily, they focused on the importance of education for gender equality and women's opportunity (Topics 4–6). Additional topics included food insecurity due to loss of education, women in science, and harassment and abuse of girls (Topics 7–10).

Figure 2: Number of tweets across the ten topics identifed using topic modeling



We took a deeper dive into six of the ten topics identified from our LDA model; Topics 2, 4, 5, and 8 were dropped from further analysis because they focused on more general activities or events rather than issues. For the remaining six topics, we analyzed the temporal trend of tweets within each topic (Figure 3), engaging in a thematic analysis of the text and corresponding hashtags pertaining to the most relevant tweets. Location data yielded confusing findings given the international nature of many of these groups and were therefore not included in the analysis.

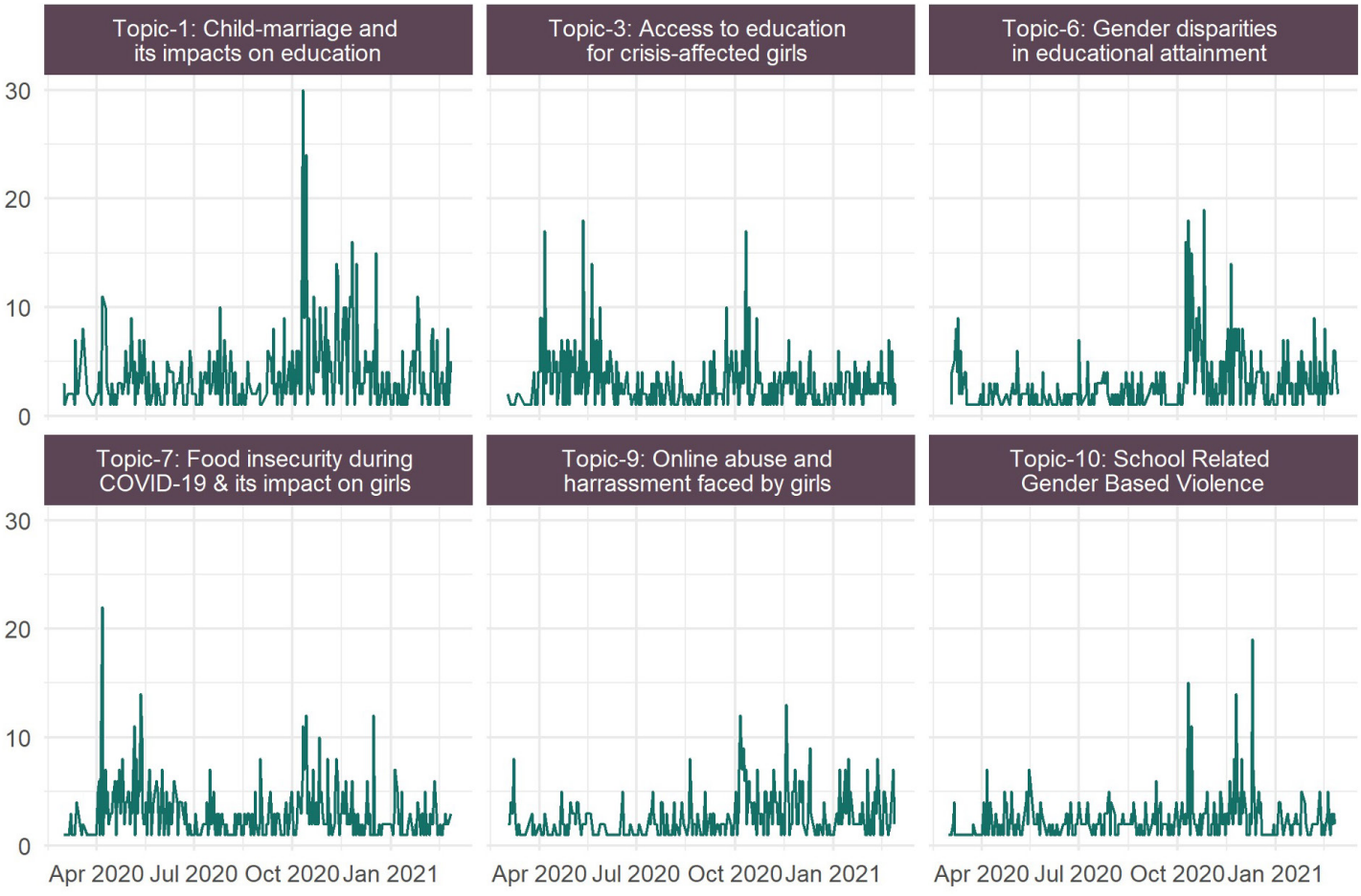
From this deep dive, we found that organizations emphasized the following topics through their tweets:

- **Child marriage and its impacts on education** as a topic highlighted how the COVID-19 pandemic has put millions of girls out of schools and is poised to force a substantial proportion of such girls into child marriage. Some of the hashtags associated with this topic were #EndChildMarriage, #GirlsFirstFund, and #ChildMarriage. Tweets in this topic peaked in October 2020, centered around the International Day of the Girl, which is celebrated on October 11.
- **Access to education for crisis-affected girls**, such as armed conflicts, forced displacement, and natural disasters, drew attention to how their educational attainment was further limited by the COVID-19

pandemic. These tweets saw multiple peaks that started in April 2020, probably when the adverse effects of COVID-19 lockdowns on education were being felt across the world. Some of the hashtags that were associated with this topic included #EducationCannotWait, #LeaveNoGirlBehind, #LearningNeverStops, and #InSolidarityWithGirls.

- **Gender disparities in educational attainment** peaked as a topic in October 2020, around the International Day of the Girl. This included tweets that highlighted the exacerbating effect of COVID-19 on the continuing disparities between boys and girls in school enrollments, sustained schooling, and school dropouts. Tweets in this topic included the hashtags of #Iamthe1stGirl and #GenerationEquality.
- **Food insecurity during COVID-19 and its impact on girls** as a topic highlighted how the pandemic led to food crises in many of low-middle income settings, which in turn affected girls' education severely. These tweets drew attention to the general scarcity of food coupled with the loss of school meals due to school closures, which lead to girls eating less, dropping out of school, and being forced into early marriage, child labor, sexual exploitation, and teenage pregnancies. This topic peaked in April 2020, possibly when some of these effects were coming to light at the onset of the pandemic.
- **Online abuse and harassment faced by girls** peaked in October and November 2020 as a topic, around the International Day of the Girl. These tweets were accompanied by the hashtag #FreeToBeOnline and highlighted the effects of such abuses on the safety and self-esteem of girls.
- **School Related Gender-Based Violence (SRGBV)** was highlighted as an impediment to girls' continued education and emphasized the need to make schools safe and create gender-responsive spaces to engender girls' educational attainment and outcomes. These tweets peaked in November and December 2020 and corresponded with the 16 days of activism against gender-based violence between November 25, 2020 and December 10, 2020. Hashtags associated with this topic included #EndSRGBV and #TogetherWeCan.

Figure 3: Number of tweets over time for topics related to gender equity in education



Implications

Our study identifies some issues around gender equity in education that were raised by organizations working in international education during the COVID-19 pandemic. We find that the issues raised on Twitter draw attention to the effects of schools being closed on girls' education. These effects include girls disproportionately dropping out of schools and an increased risk of girls being deprived of education altogether. Some of the other threats identified through our analysis include child marriage, teenage pregnancies, sexual exploitation, and gender-based violence, all of which can have a catastrophic impact on intergenerational gender equity.

We also find that organizations in this space recognized political instability, strife, and economic crises further exacerbate the effect of the pandemic on girls' education. Interestingly, the digital gender divide was not identified as a topic in our analysis as very few tweets discussed the issue. This is surprising as several studies have highlighted the digital gender divide during the pandemic as an impediment to gender equality in education.[12-14] We also find that tweets around several topics peak during significant days such as the International Day of the Girl and 16 Days of Activism, which are important ways to raise awareness and draw attention to girls' education.

With Twitter increasingly used as a tool for dissemination and communication by organizations globally, our findings can support education and gender researchers and practitioners in identifying themes that need urgent attention, particularly during the COVID-19 pandemic.

References

1. hub Ud. COVID-19 and children. 2020. https://data.unicef.org/covid-19-and-children/.

2. de Paz C, Muller M, Munoz Boudet AM, Gaddis I. Gender dimensions of the COVID-19 pandemic. World Bank, 2020.

3. Alvi M, Gupta M. Learning in times of lockdown: how Covid-19 is affecting education and food security in India. Food Security 2020; 12(4): 793-6.

4. Burzynska K, Contreras G. Gendered effects of school closures during the COVID-19 pandemic. Lancet 2020; 395(1968): 10.1016.

5. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research 2003; 3: 993-1022.

6. Tong Z, Zhang H. A text mining research based on LDA topic modelling. International Conference on Computer Science, Engineering and Information Technology; 2016; 2016. p. 201-10.

7. McCallum AK. MALLET: A Machine Learning for Language Toolkit. 2002.

8. Rehurek R, Sojka P. Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks; 2010: Citeseer; 2010.

9. Prabhakaran S. Topic Modeling with Gensim (Python). Machine Learning Plus.

10. Syed S, Spruit M. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. 2017 IEEE International conference on data science and advanced analytics (DSAA); 2017: IEEE; 2017. p. 165-74.

11. Dehingia N, Raj A. Mining Twitter Data to Identify Topics of Discussion by Indian Feminist Activists. Big Data and Gender in the Age of COVID-19: A Brief Series from UC San Diego, 2021. (accessed 03-17-2021).

12. Onyema EM, Eucheria NC, Obafemi FA, et al. Impact of Coronavirus pandemic on education. Journal of Education Practice 2020; 11(13): 108-21.

13. Hussain T. Education and COVID-19 in Nigeria: Tackling the digital divide. SOAS Blog 2020.

14. Sá MJ, Serpa S. COVID-19 and the promotion of digital competences in education. Journal of Educational Research 2020; 8(10): 4520-8.

The code for the analysis can be found here: https://github.com/akdey01/twitter_education_gender.git

The Center on Gender Equity and Health at UC San Diego's EMERGE work is supported by the Bill and Melinda Gates Foundation (Grant #s: OPP1163682 and INV-018007).