Hate Speech on Social Media in India

NBER Digitization Tutorial

Aarushi Kalra
PhD candidate in Economics, Brown University
aarushi_kalra@brown.edu, +1 401 808 7950

Education

Brown University *2018 - Present*
Ph.D., Economics
Delhi School of Economics *2015 - 2017*
Master of Arts, Economics
Lady Shri Ram College for Women *2012 - 2015*
Bachelor of Arts (with Honours), Economics

Fields of Study

Major Field: Public Economics; Minor Field: Development Economics

Course of Graduate Study

After my MA at the Delhi School of Economics (DSE), I worked as an RA at the Center for Development Economics, housed at DSE, and taught Quantitative Methods (using STATA) at the Institute of Economic Growth, New Delhi. At Brown, I completed graduate sequences in Microeconomics, Macroeconomics and Econometrics. My second year courses included Machine Learning, Market Design, Industrial Organization, Public Finance and Development Economics. I continue to participate in the Data Science for Social Good and Ambedkar-DuBois Reading Groups at Brown.

Fellowships and Awards

Graduate Program in Development Summer Funding, Brown University (2020)
The Graduate Program in Development (GPD) is an interdisciplinary initiative housed at the Watson Institute, Brown University. The program promotes research in Social Sciences in the Global South.

Stone Wealth and Income Inequality Project, Brown University (2020)
The James M. and Cathleen D. Stone Wealth and Income Inequality Project supports research and teaching on the causes and consequences of inequality.

Graduate Student Travel Fund, Brown University (2019)
Supported by Graduate School of Brown University to conduct fieldwork in India.

Faculty Prize, Lady Shri Ram College (2015)
Recognition for outstanding academic achievement and extra-curricular contribution awarded to one student from across all fifteen departments of the graduating cohort.

Exchange Scholarship, National University of Singapore (2015)
Received full scholarship to study at the Faculty of Arts and Social Sciences, NUS.

Research in Progress

1 Hate Speech on Social Media in India

This project is motivated by the recent debate on censorship of political speech on social media. The controversy sparked by President Trump's reactions to protests in Minneapolis, found Twitter and Facebook taking different positions against the President's inflammatory remarks[1]. While the question here is, should social media companies have the right to place limits to free speech in the interests of its users, we are curious about the *conditions* that give rise to online extreme speech in the first place. We believe, that such generalizable knowledge would inform this crucial policy debate, and justify the use of appropriate policy instruments.

Pohjonen and Udupa (2017) define online 'extreme speech' as digital acts that seek to polarise opinion, and that may take the form of offensive messaging targeted at individuals or communities[2]. In India, memes that target and label certain groups (including University students, journalists, and the Muslim community) as 'anti-national' are known to have incited violence. Additionally, as an increasing number of women are accessing various social media platforms, there is a renewed challenge to ensure a safe environment against spamming, and sexual harassment. While social media can grant voice to marginalized groups to express their viewpoints, negative reactions can pose harsh barriers to access these platforms, and to form networks with community members effectively (Kennedy, 2000). In India, there is an urgent need to address factors that constrain the use of digital media, especially due to reports of suicides after online harassment (Gurumurthy and Menon, 2009) and those of mob lynchings triggered my rumours on WhatsApp[3].

With more than 400 million users[4], WhatsApp has no major competitors in India. Misinformation and extreme speech received through groups and personal conversations on WhatsApp are known to alter a user's view of the world. Information flowing through the network of friends can also influence voting decisions, as noted in the case of the general election (2019) in India (Garimella and Eckles, 2020; Badrinathan, 2020). Crucially, Kamra and Williams (2019) provide evidence that digital content circulating on WhatsApp, which is characterised as hate speech informs everyday public opinion.

Given the immense popularity of WhatsApp as a communication platform, understanding flows of information on WhatsApp is key to parsing out the role of information in shaping people's attitudes. This is especially important in an atmosphere of high polarization across religious lines in India, where any WhatsApp user could have been exposed to anti-minority propaganda (Bengani, 2019) and misogynist content. This is a hard task as researchers do not have access to encrypted WhatsApp chats. As a result, previous studies have had to rely on survey data or field experiments to shed light on the kind of information WhatsApp users are exposed to and how that impacts their decisions.

We try to bridge this gap in the literature by directly observing the content that users share on WhatsApp with a novel data set from a content generation app, ShareChat. With 160 million active monthly users in India[5], ShareChat is a very popular platform where users can generate, engage with and share content in 14 non-English regional languages. ShareChat users can share content from the platform to WhatsApp directly.

It is important to address questions relating to Hate Speech and women's safety on ShareChat for two reasons. Firstly, content created on ShareChat influences the environment on WhatsApp (primarily) and Facebook (to a lesser degree), i.e. on downstream platforms where content created on ShareChat is shared. This is a claim that we also plan to verify empirically. Secondly, with an enormous user base in India, ShareChat is space where popular opinions are being formed. It is therefore, useful to understand the environment in which such posts surface and are engaged with to make sure that all marginalized groups have access to this opinion building machinery.

[1] `https://www.forbes.com/sites/chriswestfall/2020/05/30/free-speech-facebook-twitter-george-floyd-demonst`
`#5db14fa66e90`

[2] See Waseem and Hovy (2016) for a list of criterion (founded in critical race theory) to detect hate speech. Similarly, Bohra et al. (2018) construct a Hindi-English code-mixed dataset for detecting hate speech

[3] `https://www.bbc.com/news/world-asia-india-44856910`

[4] `https://techcrunch.com/2019/07/26/whatsapp-india-users-400-million/`

[5] `https://techcrunch.com/2020/09/24/indias-sharechat-raises-40-million-says-its-short-video-platform-moj-`

2 Data

ShareChat has 160 million active monthly users in India, who are creating and sharing content in 14 non-English regional languages. ShareChat can be seen as a companion service, catering to the content needs of a very large proportion of the 400 million Indian WhatsApp users. Not only will the researchers have access to observational data on the app, we will be able to conduct experimental studies, as the developers of the app have invited us to design interventions that could help make the app safer for women and other marginalized groups.

Content based social networks, such as ShareChat, are centered around topics. The interactions on ShareChat are primarily through the trending content. The data accessible to us is similar to other online social networks (Leskovec et al., 2010a,b). There are primarily two entities in the social network: content and users. We describe connections between these entities and the observations that we have access to.

There are three types of users: individuals, paid content creators, and media companies. Users can post text updates and multi-media updates: images and videos. As can be anticipated, media companies and paid content creators tend to have far more reach and content. The content created by the users is surfaced on various windows namely, trending, popular, news and so on. These windows can also be topical e.g. corona virus, government legislation, funny content, seasons' greetings, election results etc. Users can follow other users and the content generated from that connection is surfaced on the 'following' window. The data will have information pertaining to individual user i.e. name, type, and content created by the user. In addition to that, the data will have the connections from the follow feature.

A user accessing the mobile app sees content generated by other users. The user can then proceed to comment, like, download, repost, or share this content to other platforms, especially WhatsApp. We will have access to these interactions between users and content. Of these liking, reposting and commenting features are public, meaning users on ShareChat can observe the above actions on the content. The downloading and sharing features are not public, meaning that other users cannot observe who shared what on WhatsApp. The sharing feature shares the post directly to WhatsApp and the download features enables users to upload the content to other social networks, including WhatsApp. The comments under such user generated content can have nested comment threads within them. All the interactions between users are shown in the comment section of posts. The data set will also contain comments made by users on posts and nested comment threads. This essentially gives us a rich network of users and content, all of them are part of the social network graph.

All user generated content and related interactions can be varied in nature. The content can be in favour of a political party or against it. It can be in favour of or against foreign nations and marginalized groups. They could be spreading useful information or fake news. The comments can be similar in nature ranging from supportive or encouraging to hateful or abusive. We furnish examples in Appendix Figure 1. We can summarize the features of our dataset as follows:

1. **Linked to WhatsApp:** This information is not publicly available (for other users), but will be available to the researchers
2. **Enormous User Base:** With 160 million active monthly users, ShareChat is a very rich source of data. For comparison, WhatsApp has 400 million users in India
3. **Localized engagement with content:** Content is highly localized by geographic location, local language, and is geo-tagged
4. **Mapping Networks:** The researchers can map user to user, user to content, and content to content links
5. **Private Beliefs:** We can distinguish between users social and private beliefs by noting different kinds of engagement activity on the app. 'Sharing' behaviour is not visible to other users but 'Commenting' behaviour is. Researchers observe both.
6. **Representation:** of a specific socio-economic group, seemingly unrepresented on Twitter and Facebook

References

Badrinathan, S. (2020). Educative interventions to combat misinformation: Evidence from a field experiment in india.

Bengani, P. (2019). India had its first 'whatsapp election.' we have a million messages from it. *Columbia Journalism Review*.

Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., and Shrivastava, M. (2018). A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pages 36–41.

Garimella, K. and Eckles, D. (2020). Images and misinformation in political groups: Evidence from whatsapp in india. *arXiv preprint arXiv:2005.09784*.

Gurumurthy, A. and Menon, N. (2009). Violence against women via cyberspace. *Economic and Political Weekly*, pages 19–21.

Kamra, L. and Williams, P. (2019). Strategies to tackle extreme speech on whatsapp must bring together socio-political, digital worlds. *Scroll.in*.

Kennedy, T. L. (2000). An exploratory study of feminist experiences in cyberspace. *CyberPsychology & Behavior*, 3(5):707–719.

Leskovec, J., Huttenlocher, D. P., and Kleinberg, J. M. (2010a). Governance in social media: A case study of the wikipedia promotion process. *ArXiv*, abs/1004.3547.

Leskovec, J., Huttenlocher, D. P., and Kleinberg, J. M. (2010b). Signed networks in social media. In *CHI*.

Pohjonen, M. and Udupa, S. (2017). Extreme speech online: An anthropological critique of hate speech debates. *International Journal of Communication*, 11:19.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

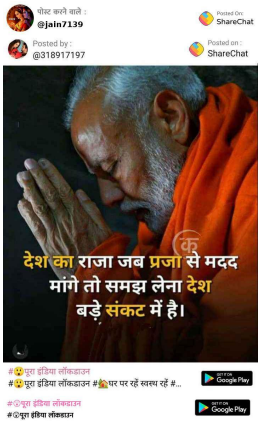
Appendix



(a) Hate speech towards China



(b) Message approving Police Violence during lockdown



(c) In support of the Indian Prime Minister



(d) Suggesting Modi consult a religious Guru who can end the pandemic



(e) Depicting the plight of the unemployed in India due to the pandemic

Figure 1: Examples of user generated content on ShareChat