9th International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

Characterizing networks of hate on Nigerian Twitter

Keywords: networks, hate speech, demographic inference, Twitter, Nigeria

Extended Abstract

The age of social media promised an era of connectivity, diversity and inclusion. This technology, however, has had unintended consequences, including the amplification of extreme voices in ways that were not possible with traditional media [1]. In particular, online hate speech is thriving on social media and has been associated with outbursts of physical violence and hate crimes [2]. Despite its serious potential consequences, online hate speech's dynamics still elude us. Most past research on the topic developed tools to detect hate on small curated datasets [3]. Prior work in this space also leveraged these tools to characterize online hate speech at the user [4] and the conversation [5] level. Yet, most of this work has been conducted on a small scale, limiting the validity of its findings. Additionally, this research has been primarily focused on the Global North, despite the alleged role of social media in violent events in the Global South, such as the Rohyngia massacre[1]. Consequently, the prevalence, scope and dynamics of hate speech on social media in the Global South remain unclear.

In this ongoing work, we aim to bridge this gap and develop a methodological framework to study hate speech on social media at the population level in the Global South. We focus on Nigerian Twitter as it provides an opportunity to study online hate speech at the highest level. Exemplifying the issue, Twitter was banned by the Nigerian government in June 2021, supposedly fallout from the platform's deletion of a tweet by President Buhari in which he incited violence towards the Biafran separatists. We gathered a dataset of 1 billion tweets posted between January 2009 and February 2023, corresponding to the full timelines of 1.8 million Nigerian users whose profile location is in Nigeria. Recent estimates[2] suggest that Nigeria has about 3 million Twitter users, which indicates that our dataset provides good coverage of the Nigerian Twitter population.

To better understand the profile of hate producers, we first created a pipeline to infer the ethnicity, gender, and religion of users. For evaluation purposes, we labeled a sample of 2,000 Nigerian Twitter users. We also collected 2,500 first names labeled with each demographic attribute to perform a first inference layer supported by name matching. While robust, this strategy only covers cases where a labeled name is found in a user's display name. For the remaining users, we leveraged the followership network using label propagation and achieved full population coverage with high accuracy. Further, we inspected the demographic composition of the whole user population using our labeled random sample (Figure 1). We found that the Christian South is over-represented, and the Muslim North is under-represented among users. We also found that men are over-represented on Twitter and that users are almost exclusively located in large urban areas, as previously observed on the US Twitter population [6].

Further, we leveraged transfer learning to detect hate speech in Twitter discourse. We first built lists of community names and hate words specific to the Nigerian context and sampled tweets containing each possible combination of hate word and community name. We then

[1] https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/

[2] https://africacheck.org/fact-checks/reports/forty-million-twitter-users-nigeria-how-pollsters-flawed-figure-became-fact

labeled this tweet sample in three mutually exclusive categories, namely “neutral”, “offensive” and “hateful”. We further sampled and labeled tweets using active learning and achieve an average precision above 90%. Beyond the hatefulness of a tweet, we further labeled the targeted community of each sampled tweet labeled as hateful. We used supervised learning to accurately predict each hateful tweet’s targeted community. For our preliminary analysis, we focus on ethnic and regional communities and bundle these into three distinct groups: “North”, “South” and “East”.

We run our demographic inference and hate detection models on the full population, creating a labeled dataset at the user level. By combining hate speech classification and the user demographic inference, this approach results in a nuanced index that tracks not only general and community volume (Figure 2) but also the flow of toxic content between communities (Figure 3). We find that, independently of the ethnicity of the hateful user, most of the hate content is directed towards northern Nigeria. We further analyze the homophily between users flagged as hateful and those they have interacted with and report the results in Figure 4. For each hate target community C, we find that hateful users towards C have a significantly (Mann-Whitney U p-value $< 10^{-6}$) larger fraction of hateful neighbors towards C in the retweet network, suggesting homophily in the flow of hateful content.

Next, we will analyze the characteristics of hate speech on Nigerian Twitter at three levels: user, conversation, and tweet. From the user perspective, we will quantify how central hateful users are by computing centrality measures in the interaction networks controlling for demographic attributes. At the conversation level, we will investigate what type of content triggers hate. We will also compute the probability of a hateful reply depending on the original post and social relationship between producer and replier (dyad-level), as well as the relationship between the density of the followers’ network of users taking part in a conversation and the overall conversation hatefulness (group-level). Finally, at the tweet level, we will look closer at the temporal evolution of hate (Figure 2). We will also characterize the spread and reach of hateful tweets and compare them with non-hateful content. This will indicate whether emotionally charged content is more likely to percolate through the conversation.

References

[1] Chris Bail. *Breaking the social media prism: How to make our platforms less polarizing*. Princeton University Press, 2022.

[2] Karsten Müller and Carlo Schwarz. “Fanning the Flames of Hate: Social Media and Hate Crime”. In: *Journal of the European Economic Association* 19 (Aug. 2021), pp. 2131–2167.

[3] Thomas Davidson et al. “Automated Hate Speech Detection and the Problem of Offensive Language”. In: *Proceedings of the International AAAI Conference on Web and Social Media* 11 (2017). Number: 1, pp. 512–515.

[4] Manoel Ribeiro et al. “Characterizing and detecting hateful users on twitter”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 12. 1. 2018.

[5] Martin Saveski, Brandon Roy, and Deb Roy. “The structure of toxic conversations on twitter”. In: *Proceedings of the Web Conference 2021*. 2021, pp. 1086–1097.

[6] Alan Mislove et al. “Understanding the demographics of Twitter users”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 5. 1. 2011, pp. 554–557.

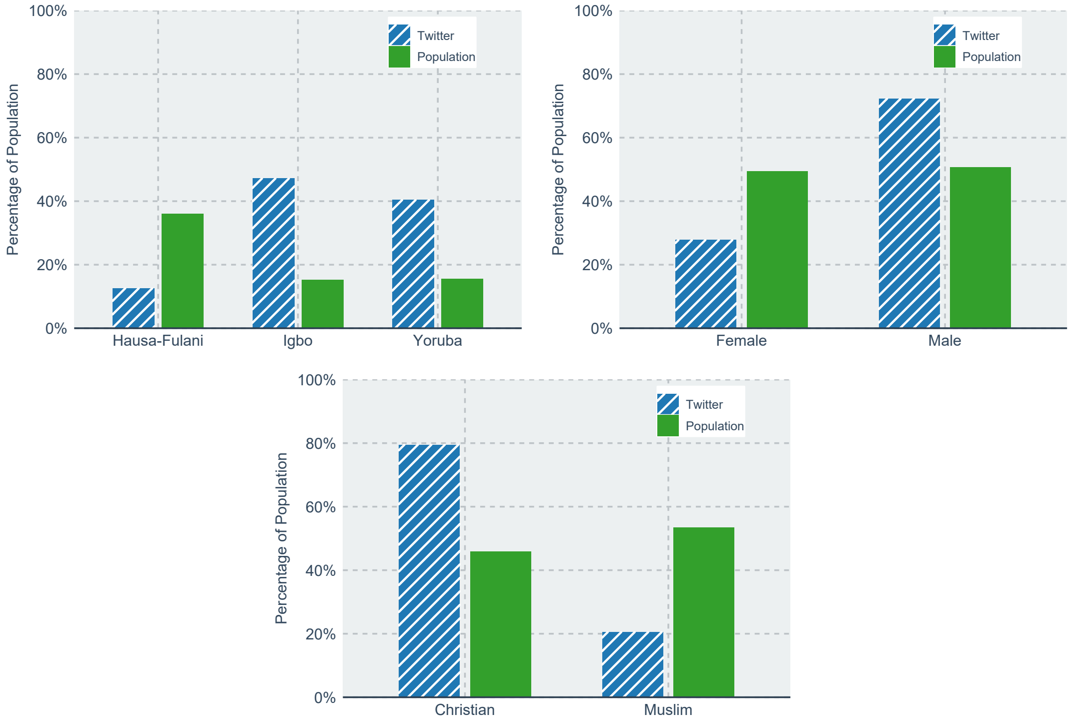

Figure 1: Demographic attributes comparison between Twitter users (blue) and the overall Nigerian population (green) for ethnicity (top-left) gender (top-right) and religion (bottom).

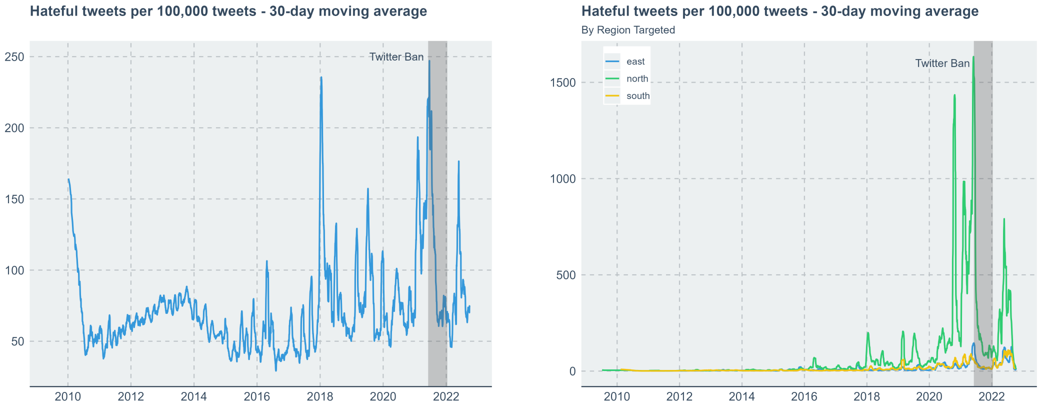

Figure 2: Trends in volume of hateful tweets targeting all groups (left) and targeting groups from the East, North, and South as indicated by the legend (right).

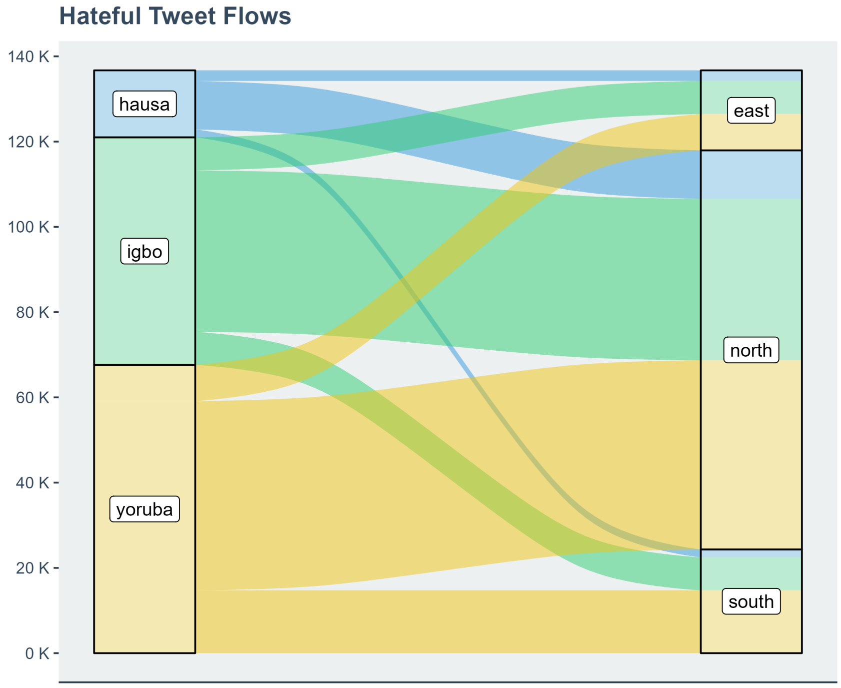

Figure 3: Flow of hateful content emanating from major ethnic groups (left) and targeting regional groups (right).

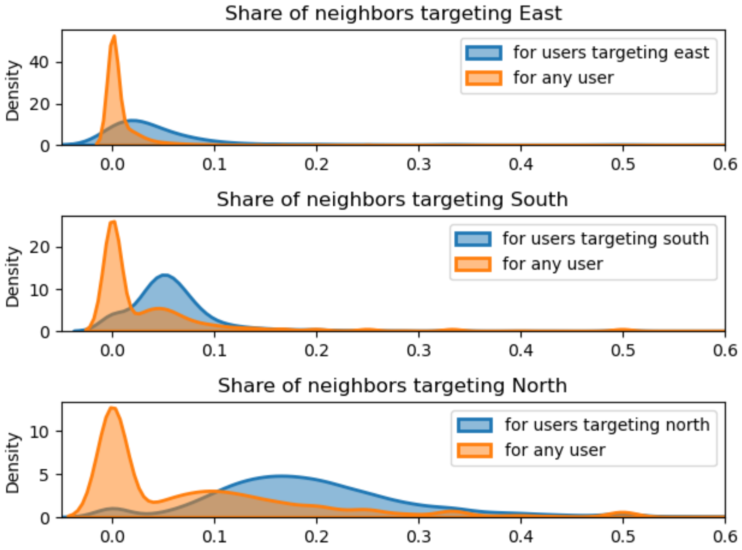

Figure 4: Kernel density estimations of the proportion of neighbors hateful towards target community C – where C = East (top), C = South (middle), and C = North (bottom) – for users hateful towards C (blue) and any user of the population (orange).