Methods Supplement for "Ethiopia Twitter Intervention"

Megan A. Brown

January 10, 2022

1 Data Collection

We collected tweets related to the crisis in Ethiopia from November 1, 2021 to November 8, 2021 using Twitter's Search API. We collected pro-government tweets by matching tweets containing the hashtags "#unityforethiopia," "#ethiopiaprevails," "#tplfterrorists," "#tplfisaterroristgroup," "#tplfterroristgroup," "#tplfiswarcriminal," "#handsoffethiopia," "#tplfsurrendernow," "#nonegotiationwithtplf," "#tplfisthecause," "#nomoretplf," and "#nomore" during the time period of interest. We collected pro-Tigray tweets containing the hashtags "#tigraygenocide," "#stopwarontigray," "#stopthewarontigray," "#standwithtigray," "#istandwithtigray," "#tigrayfamine," "#tigraymassarrests," "#tigray," "#1yearoftigraygenocide," "#callitagenocide," "#tigrayans," "#tigrayshallprevail," "#endtigraysiege," "#stopbombingtigray," "#stoptigrayfamine," "#reconnecttigray," and "#opentigray" during the time period of interest.[1] Note, these matches for hashtags are not case sensitive, meaning a tweet containing '#nomore' would be collected, as would a tweet containing '#NoMore.'

For each tweet in our dataset, we use Perspective, an open-source API by Jigsaw and Google's Counter Abuse Technology team, to label the tweets. Perspective uses machine learning to identify abusive content of six different types: toxicity, insult, identity attack, profanity, threat, and sexually explicit.[2] For the purposes of this analysis, we are interested in toxicity and threat content for each of the tweets.[3] We analyze each of these tweets using Perspective, resulting in "scores" for each of the quantities of interest. A "score" is a value between zero and one (a probability, so to speak), generated by Perspective, such that scores closer to zero indicate that the message does not contain the value of interest (e.g. a low score for toxicity would indicate that the comment is unlikely to contain toxic content while a high score would indicate that the comment is likely to contain toxic content).

[1] Hashtags were sourced from the Media Manipulation Casebook Report regarding disinformation campaigns related to the conflict: `https://mediamanipulation.org/case-studies/dueling-information-campaigns-war-over-narrative-tigray`. The original set of hashtags in this report were augmented to include more recent hashtags that commonly co-occured with the hashtags in the report.

[2] Full data definitions can be found here: `https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages`

[3] Importantly, Perspective only allows for analysis in languages for which they have developed classifiers. In this piece, we are able to analyze English-language tweets about the conflict. However, further analysis should analyze these trends in local languages.

2 Data Aggregation

For pro-government tweets and pro-Tigray tweets, we calculate the number of original tweets per hour in that category. We take the 24-hour rolling average. In Figure 1 we show the number of tweets per hour that contain pro-government hashtags.

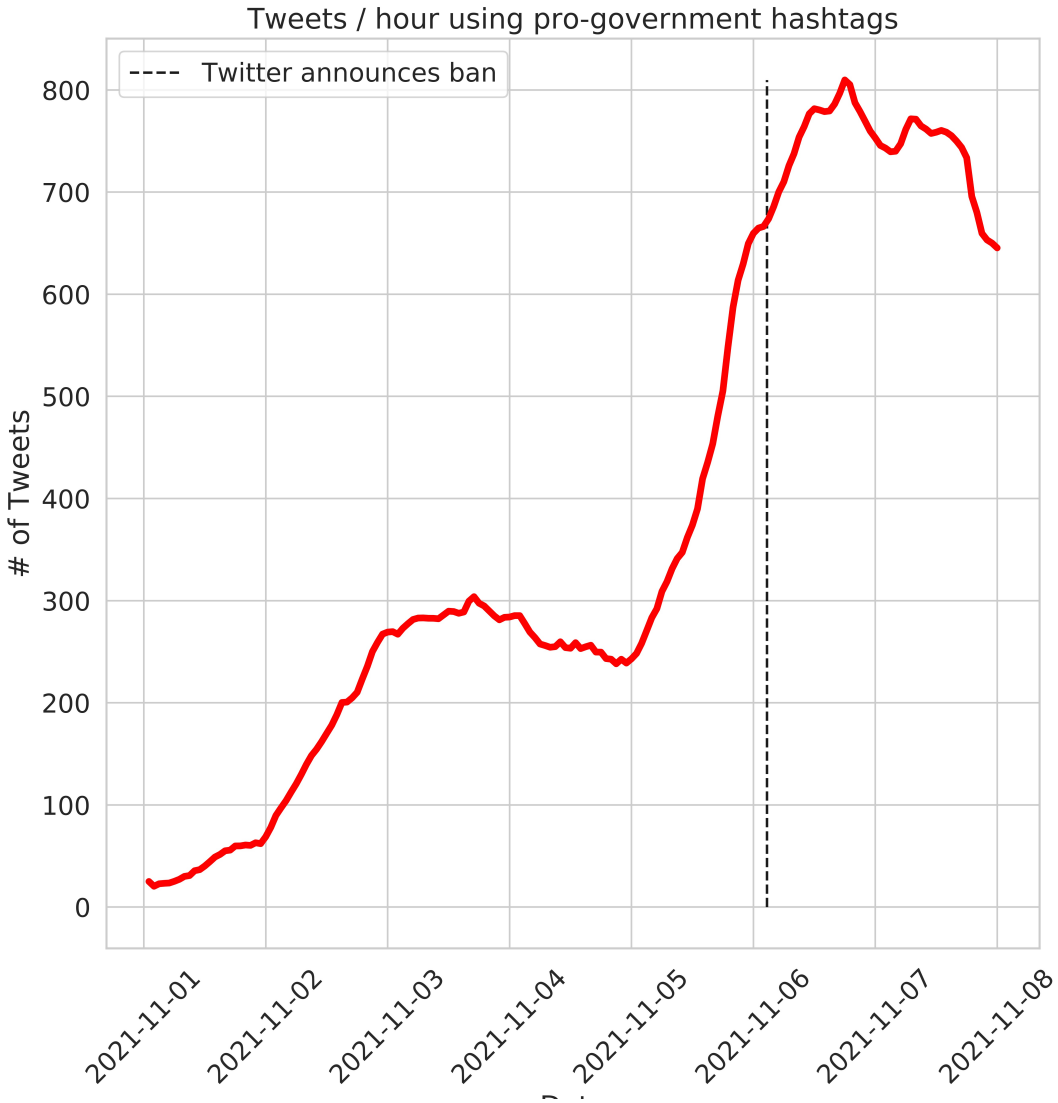

Figure 1: 24-hour rolling average of number of tweets containing pro-government hashtags. We mark the date Twitter announced their intervention in trending topics with a dashed black dashed line. Figure: Megan A. Brown, NYU Center for Social Media and Politics

In Figure 2 we show the number of tweets per hour that contain pro-Tigray hashtags.

Finally, we calculate the average proportion of tweets containing toxic and threatening language per hour and calculate the 24-hour rolling average, separated by pro-government and pro-Tigray tweets. In Figure 3 we show the proportion of tweets containing toxic and threatening language during the period of interest.

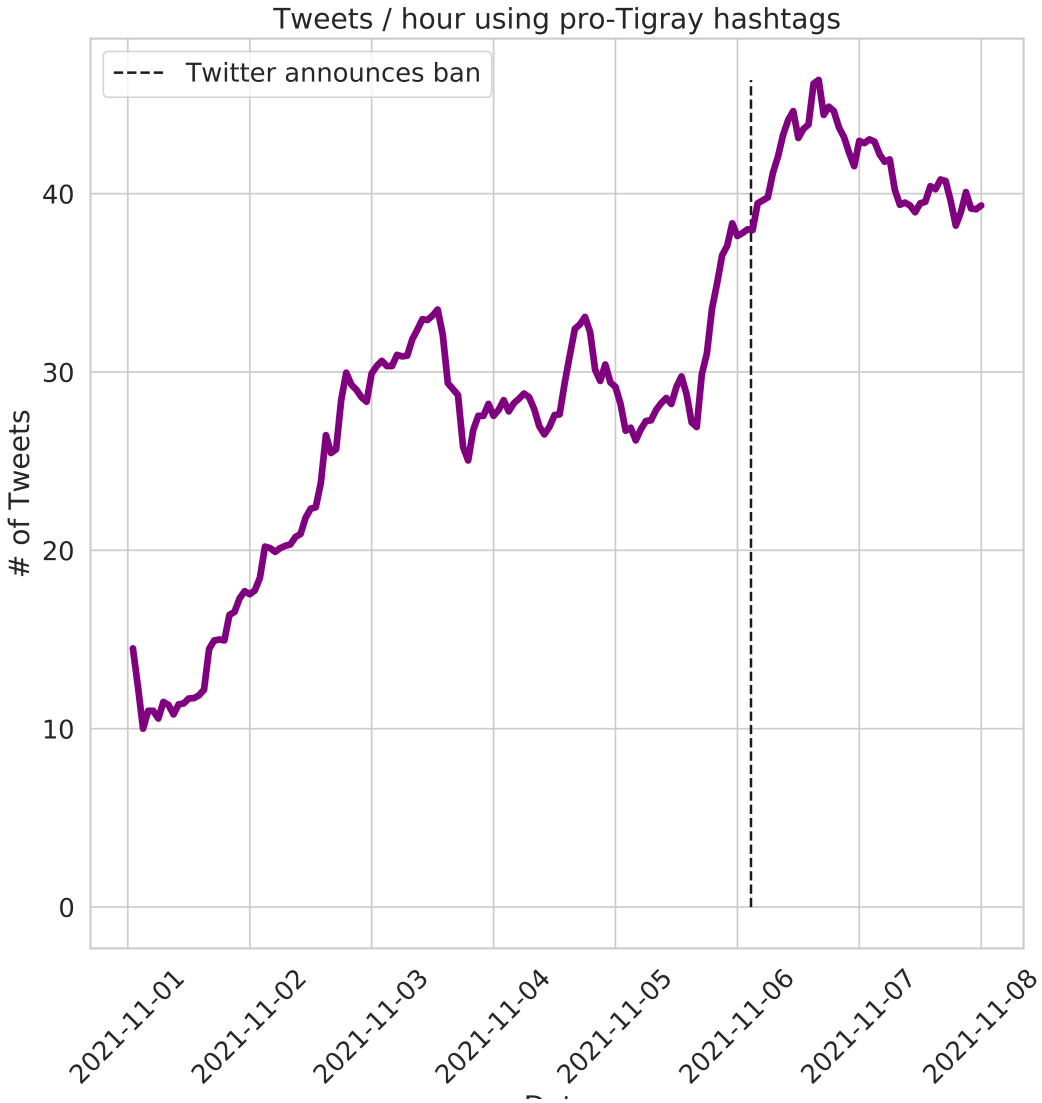

Figure 2: 24-hour rolling average of number of tweets containing pro-Tigray hashtags. We mark the date Twitter announced their intervention in trending topics with a dashed black dashed line. Figure: Megan A. Brown, NYU Center for Social Media and Politics

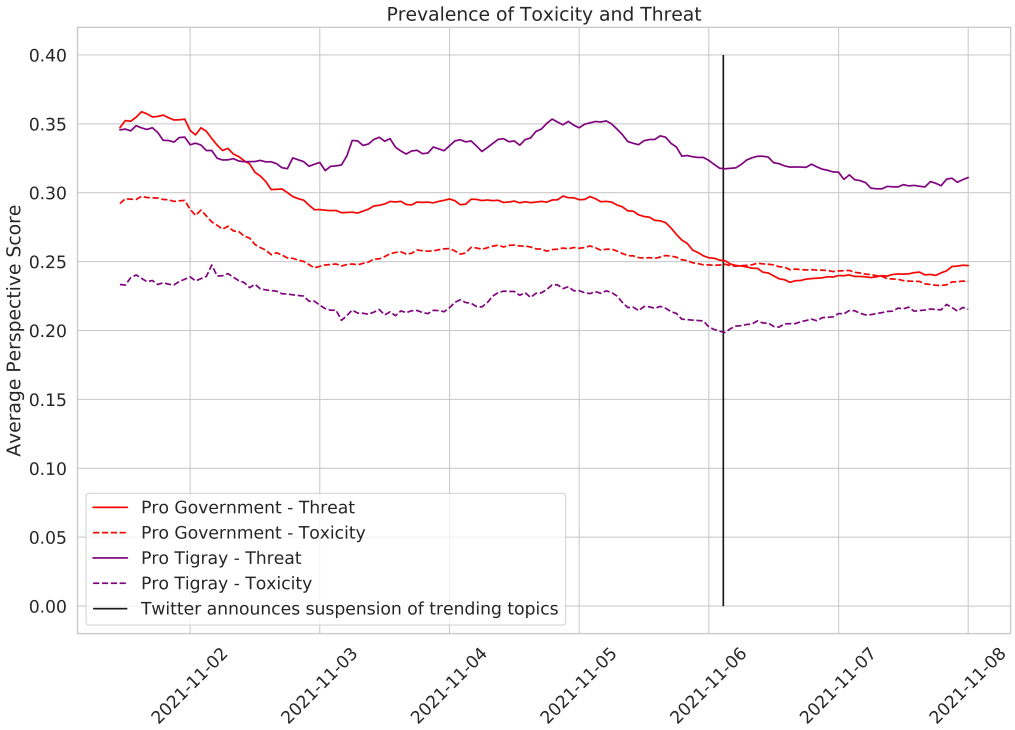

Figure 3: 24-hour rolling average of toxic (dashed) and threatening (solid) language for tweets containing pro-government (red) and pro-Tigray (purple) hashtags. We mark the date Twitter announced their intervention in trending topics with a solid black line. Figure: Megan A. Brown, NYU Center for Social Media and Politics