README for Kingdom of Trolls? Influence Operations in the Saudi Twittersphere

Christopher Barrie and Alexandra A. Siegel

3/24/2021

File structure

The below describes the overall file structure:

```
#get tweets from Saudi Arabia
  0_00_saudistream.ipynb
#generate .csv of tweets from Saudi Arabia
  0_01_saudistream_gencsv.ipynb
#get last 3,200 tweets of users in Saudi Arabia
  0_02_get_saudiusertweets.R
#combine all Saudi Arabia tweets (SA sample)
  0_03_combine_saudiusertweets.R
#get IO tweets from Twitter releases and filter to Saudi Arabia
  0_04_get_saudi_IO_tweets.R
#get most followed Saudi users from socialbakers list
  0_05_get_most_followed_SAusers.R
#read and clean TOP and NEWS Saudi users  from Historical Power Track stream
  0_06_clean_SA_top_accounts.R
#read Saudi geolocated tweets from Historical Power Track stream
  0_07_get_SA_geo_tweets.R
#plot geolocated users clipped to Saudi Arabia borders
  0_08_plot_SA_geo_tweets.R
#get last 3,200 tweets of all users geolocated to Saudi Arabia
  0_09_get_geo_user_tweets.R
#combine all geolocated tweets for GEO sample
  0_10_combine_geo_user_tweets.R


#filter IO tweet releases and hydrate tweet IDs for SA, GEO, and NEWS samples
  1_00_hydrate_analysis_data.R
#compare mentions of IO accounts across SA and GEO samples
  1_01_compare_inf.R
#compare engagement with IO accounts to SA, GEO, NEWS, and TOP
  1_02_compare_engagement.R
#compare contentious topic engagement for IO accounts against SA, GEO, NEWS, and TOP
  1_03_compare_topics.R
#get account names of users mentioning IO account
  1_04_get_mentioning_accounts.R
#generate matrix for network visualization (done in Gephi)
  1_05_network_analysis.R
#get top users for coding by account type
```

```
  1_06_top_users_for_human_coding.R
#get top hashtags shared by news accounts
  1_07_top_hashtags_news_tweets.R
#compare SA and GEO sample characteristics
  1_08_compare_samp_characteristics.R
#compare engagement with other political across IO, SA, GEO, NEWS, and TOP
  1_09_compare_alt_topics.R

#folder structure for relative file paths in scripts:
  data
      analysis
      output
          plots
          saudigeousertweets
          saudigeousertweets_combined
          saudiusertweets
      replication_tweetIDs
          GEOusertweets19_allIDS.txt
          IOtweets19_allIDS.txt
          NEWSusertweets19_allIDS.txt
          SAusertweets19_allIDS.txt
          TOPusertweets19_allIDS.txt
      shapefiles
           SAgeopoints.dbf
           SAgeopoints.prj
           SAgeopoints.shp
           SAgeopoints.shx
           SAgeopoints_clipped.dbf
           SAgeopoints_clipped.prj
           SAgeopoints_clipped.shp
           SAgeopoints_clipped.shx
           sau_adm0
               SAU_adm0.cpg
               SAU_adm0.csv
               SAU_adm0.dbf
               SAU_adm0.prj
               SAU_adm0.shp
               SAU_adm0.shx
```

Data collection and hydration scripts

Scripts numbered with prefix "0_" are used to collect the tweet datasets used in the analyses. The TOP and NEWS tweets were streamed with the proprietary Historical Power Track API and so we do not include code used for their collection. The original GEO tweets were also collected using the Historical Power Track API and so we do not include code used for their collection. The IO tweets were taken from the unhashed versions, which are accessible via application to Twitter Transparency here. The publicly available unhashed versions, available at the same address, are an alternative to the unhashed data: they differ only in that the usernames and other identifying information of users with < 5000 followers are omitted.

The scripts numbered with prefix "0_"are included in these replication files for the sake of completeness and transparency. Given that thecdata were collected in 2020, and the API returns only a random subsample of tweets, using the same scripts now to query the Twitter API will not return the same data.

As such, we provide an additional script "1_00_hydrate_analysis_data.R" that can be used to hydrate tweets IDs for all non-IO datasets; i.e., the SA, GEO, NEWS, and TOP samples. It is worth noting that hydrating these tweets will not result in exactly the same sample sizes or tweets as tweets that have subsequently been deleted will not be retrievable. We also provide the IDs for the IO tweets, which will save researchers time if they wish to filter the Saudi Arabia IO tweets provided by Twitter Transparency to include only those tweets we include in our analyses. Example scripts for how to do this are also provided.

Analysis scripts

All subsequent scripts with prefix "1_" include the code used to analyze the five tweet samples used in the article. The purpose of each script is describe in the file tree above.

Sample characteristics

The sample characteristics for our final five analysis samples are tabulated below.

Sample	N. tweets	N. accounts
IO	9,826,132	4,536
SA	4,697,930	7,418
GEO	11,332,330	20,115
NEWS	449,320	15
TOP	145,243	41