Computational methods for querying and sampling the Twitter disinformation datasets

Holmberg, Nils

2020

Computational methods for querying and sampling the Twitter disinformation datasets

Nils Holmberg[1,*]

2020-11-13

Abstract

This study attempts to apply computational methods to the Twitter Election Integrity Datasets in order to derive a basic descriptive overview of this disinformation data, and to suggest some possible routes for developing these methods to address future research questions. The results indicate substantial variations in tweet frequency over time and geographical regions, as well as differences in relative importance of tweet words across regions. Aggregated tweet measures provide basic descriptive statistics for the datasets.

[1] Department of Strategic Communication, Lund University

[*] Correspondence: Nils Holmberg <nils.holmberg@isk.lu.se>

1 Introduction

The 2016 US election was followed by an intense debate concerning Russian influence operations on social media. Partly in response to this debate, social media platform Twitter took an initiative in 2018 to release the Twitter Election Integrity Datasets (TEID). The intention of these data releases has been to provide transparency concerning the countermeasure of taking down the identified accounts, and to provide communication researchers with a somewhat objectively verified database of social media disinformation content. The present research aims at providing a descriptive overview of the entire TEID material currently available, and also outline some computational methods for summarizing key statistics about the data, as well as obtaining random samples of tweet texts from the database. In order to achieve this, the following research questions have been formulated:

- **RQ1:** What computational methods can be used to query and sample the TEID material?
- **RQ2:** How are disinfo tweets distributed over time and across geographical regions?
- **RQ3:** How are tweets distributed across languages, specifically English and Swedish?
- **RQ4:** What are the most important words in tweet texts across regions and languages?

The overarching aim of the RQs presented above is to suggest approaches through which TEID big data can become more accessible to a larger community of researchers. Since this data consist of about 210 million tweets across multiple text archives, it was necessary to implement methods for quickly searching through archives in a distributed and computationally effective way.

```
# import packages
library(parallel)
library(tidyverse)

# tweet archive search function
tweets_extract <- function(tweet_text, keyword) {
    if (str_detect(tweet_text, keyword)) {
        return(tweet_text)
    }
}

# distributed keyword search
results <- parallel::mclapply(archive, tweets_extract, keyword="keyword", mc.cores=20)
```

To solve this challenge, the "tidyverse" package in the R language was used to efficiently scan through large amounts of Twitter text data, while the "parallel" package was used to distribute the process over multiple cores. The pseudo-code above was submitted as a SLURM batch job at the LUNARC High-Performance Computing cluster at Lund University (project SNIC 2019/6-71).

2 Methods

The methods described in this section provide a tentative answer to RQ1, and the workflow aims at being as reproducible as possible, so as to allow for future improvements by other researchers.

2.1 Datasets

The complete list of 31 tweet datasets provided on the Twitter website was downloaded in zip format (e.g. "ira_tweets_csv_hashed.zip"). Archives containing media and account information were ignored. The downloaded zip archives were then decompressed into 83 csv files. The text files were renamed according to 14 geographical regions (e.g. Saudi Arabia, Egypt, and the UAE was collectively renamed as "SA_EG_AE", cf. Results, Table 1, below). Individual csv files larger than 1.8G were split into three parts to simplify further processing. The final step of preparing the dataset for analysis consisted in normalizing the tweet csv files such that each row constituted one tweet, that all values were tab-separated, and ensuring that tweet IDs were unique across the entire dataset.

2.2 Analyses

The first step of the data analysis consisted in aggregating the normalized csv files in order to obtain basic descriptive statistics, such as the number of disinfo tweets in each geographical region (RQ2). All tweets defined as English were extracted, and out of this subset, those tweets that explicitly mentioned Sweden were extracted (RQ3). A random sample of 1000 tweets was then drawn from each of the 83 csv files, resulting in a total of ca 80K English tweets, and ca 13K English tweets referring to Sweden. After cleaning the text data from any non-ASCII characters, ca 60K tweets remained, and this sample was used to analyze important words in tweet texts using the TF-IDF metric (RQ4). Aggregated and sampled data have been published on Kaggle (https://www.kaggle.com/nilsholmberg/scom-teid).

2.3 File formats

A possible alternative to delivering the TEID datasets as csv files, would be to utilize the json format. This would probably entail a slightly larger overhead in terms of file sizes, but would provide easier parsing of the data, without requiring normalization procedures.

3 Results

Tweets by time and region in response to RQ2 (**Figure 1**). Red dots indicate English language tweets, while green dots indicate English tweets explicitly referring to Sweden (N=60K sample).

The aggregated data in reponse to RQ3 (**Table 1**) shows tweet counts by region, user counts, average proportion of English tweets, and average number of tweets per user (N=210M).

Table 1: Aggregated tweets data.

region	t_count	t_count_en	t_count_sv	u_count	t_prop_en	t_per_u
bangladesh	26212	26212	4	11	1.0000000	2382.9091
catalonia	9489	2043	0	76	0.2153019	124.8553
china	14196324	7419013	2564	29639	0.5226010	478.9745
ecuador	700240	89003	91	787	0.1271036	889.7586
ghana	39964	39964	22	60	1.0000000	666.0667
honduras	1165019	290612	32	3007	0.2494483	387.4357
indonesia	2700296	1603233	1601	716	0.5937249	3771.3631
iran	9314829	7007637	4634	9827	0.7523098	947.8812
russia	13124186	5764162	4750	4856	0.4392015	2702.6742

region	t_count	t_count_en	t_count_sv	u_count	t_prop_en	t_per_u
sa_eg_ae	78054349	18061858	6318	35696	0.2314010	2186.6413
serbia	43067074	18825548	4217	42160	0.4371216	1021.5150
spain	56712	20643	0	216	0.3639970	262.5556
turkey	36948537	3041700	4487	22520	0.0823226	1640.6988
venezuela	10504995	1655247	1576	2475	0.1575676	4244.4424
total	209908226	63846875	30296	152046	0.4408643	1550.5551

The most important words in tweet text according to TF-IDF weighting in response to RQ4 (**Figure 2**). Comparison between four prolific disinfo regions (China, Iran, Russia, SA_EG_AE, N=60K sample).

4 Discussion

The present investigation shows that computational methods can be applied to derive basic descriptive measures related to the entire TEID dataset.

4.1 Tweets by time and region

The tweet by time and region analysis indicate substantial variations in frequency over time. Such analyses could be developed to study disinformation campaigns leading up to important events such as elections. Analyzing patterns in tweet frequency and content preceeding regional events could provide indicators of disinformation intent, especially in combination with various types of social network analyses for identifying coordinated account clusters.

4.2 Aggregated tweet data

The aggregated data provided a descriptive overview of the entire 210M tweet TEID dataset. Particularly, this analysis showcases how the data can be subset by filtering out english language tweets, and searching for tweets mentioning Sweden. This approach could be further developed to investigate the prevalence of specific disinformation phenomena, for example brand hijacking (e.g. by searching for the IKEA brand). It should be noted that the aggregated data analysis produced some mismatch in number of users compared to the official metadata delivered by Twitter. Such issues need to be resolved, e.g. by providing instructions on how to correctly read and parse the files.

4.3 Tweet word importance

Finally, the analyses of important words in disinformation tweets across regions could be used to design effective countermeasures that take into account how particular themes resonate with target audiences. Word importance within four geographical regions was measured using the widely utilized "term frequency, inverse document frequency" measure (TF-IDF). By indicating which types of words and themes are more prominent in which disinformation regions, the analysis could potentially be used as one of several signals that contribute in identifying disingenuous content.

4.4 Future research

An interesting avenue for future research would be to employ machine learning techniques that use tweet text content in order to automatically predict tweets that have a high likelihood of orginiating from disinformation campaigns. This approach would require access to similar amounts of "normal" twitter content generated by authentic users (e.g. https://archive.org/download/archiveteam-twitter-stream-2018-10/). By combinig these data, it seems feasible to train deep learning architectures such as tensorflow to learn the relationship between text content and a binomial outcome variable, i.e. disinformation or normal. Another scenario that should be investigated in future research is to combine content analyses of Twitter posts with media effects research, e.g. eye movements (Holmberg 2016), to study physiological effects of disinformation content on users compared to normal content, along with the effects of flagging suspicious twitter content on human perception.

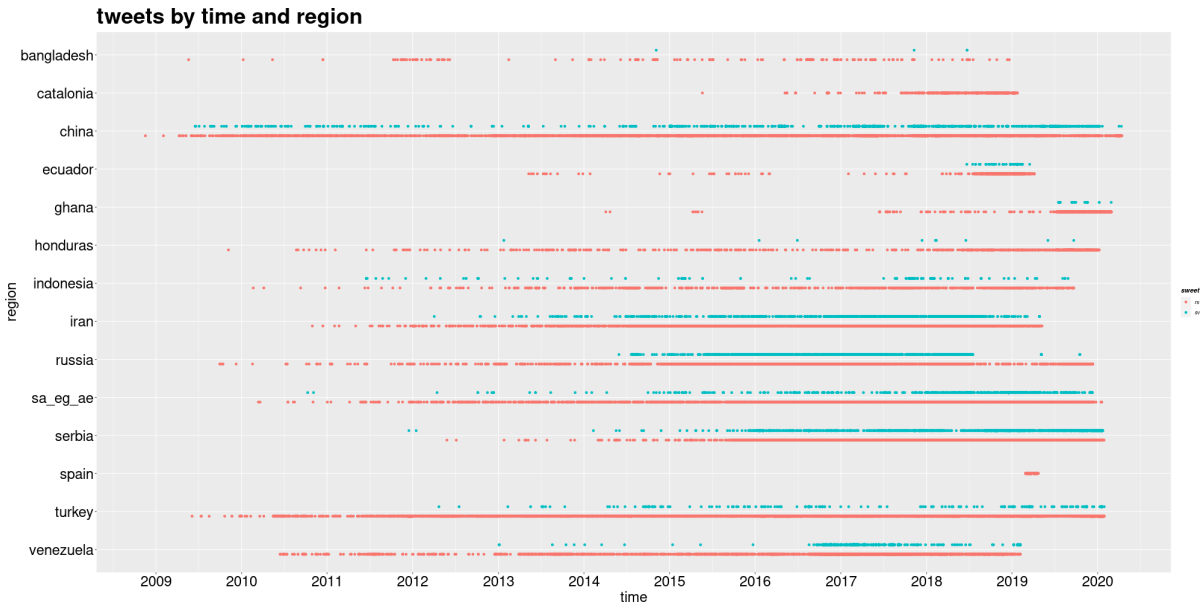

Figure 1: Tweets by time and region.

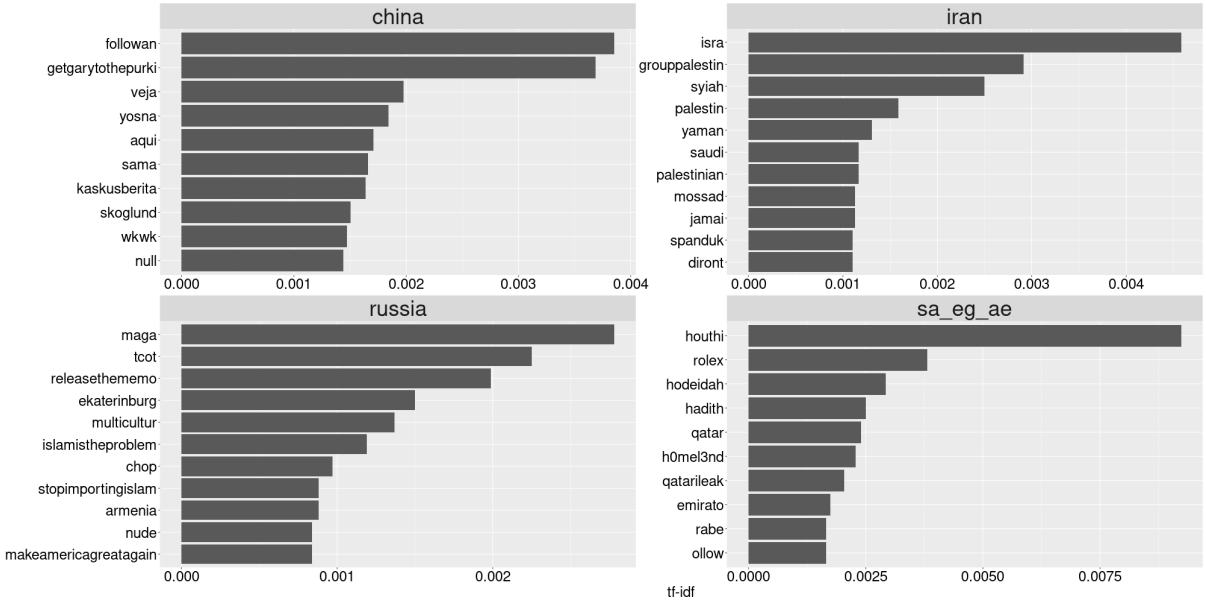

Figure 2: Most important words in tweet text.

5 Acknowledgements

The current project (SNIC 2019/6-71) was facilitated by the LUNARC computing cluster at Lund University.

References

Holmberg, Nils. 2016. *Effects of Online Advertising on Children's Visual Attention and Task Performance During Free and Goal-Directed Internet Use: A Media Psychology Approach to Children's Website Interaction and Advert Distraction.* Lund University.