Twitter Election Monitor Methodology

Jian Cao, Nicholas Adams-Cohen, and R. Michael Alvarez*

October 13, 2020

Since 2014, our research has collected Twitter data that tracks conversations about election administration and voting technology. We now collect the Twitter data, and post visualizations of the data in real-time on our Monitoring the Elections website (currently the monitor is available at https://rmichaelalvarez.github.io/twitter-monitors.html). We have argued elsewhere that this crowd-sourcing of election monitoring is an effective way to study election successes and problems in real-time, which is of particular importance in the 2020 general election.[1]

As there is a great deal of interest in this election and attention to the details of election administration, while at the same time the COVID-19 pandemic has made it difficult to conduct in-person election observation, we argue that work like this is generating important and valuable data for examining the integrity of this election. That is, the crowd-sourcing of election monitoring can provide important real-time data about election-related issues as they arise during this pandemic.

The Twitter monitor collect all tweets that mention the election keywords (listed below), categorizes and geo-locates them, then pushes the analytics to the website in real-time.

The Twitter monitor has three parts: data collection, data processing & storage, and analytics. We use our long-term social media monitoring architecture (https://arxiv.org/abs/2005.02442) to collect real-time election tweets. The architecture is deployed on the Google Cloud Platform (GCP), and requests tweets from the Twitter Streaming API, filters them with keywords, and publishes them in a Pub/Sub topic. The GCP streaming service Pub/Sub ensures the architecture can ingest and unload thousands of tweets in seconds and provides extra buffer storage to improve the system's failure tolerance. After passing through the Pub/Sub channel, the tweets are saved in a Cloud Storage bucket for subsequent analysis.

To study what election issues the tweets contain, and where these conversations took place, we then categorize and geo-locate the stored tweets. The Twitter monitor loads the collected tweets from a Cloud Storage bucket and classifies them into eight categories based on the text in the tweet. We then geo-locates the tweets that have sufficient location information, and associates those with state codes. We store the classification and geo-location information in a MariaDB database for the website to access.

We use a GCP server to read the MariaDB database and update the analytics on the website every 15 minutes. There are three types of analytics: hourly frequencies of the election topics, daily composition, and geo-distribution within each category. The hourly frequencies figure shows the counts of tweets that belong to the categories and how they change over time. The figure only covers a day, but a reader can switch to the past days using the timeline below the figure (the timelines are available for all visualizations). The daily totals figure compares the categories in a day. The state maps show, for each category, where the conversation most likely took place. The darker red states tend to have a higher number of tweets per 1,000 capita, which means the election topic is trending in these states.

Keywords:

*Cao and Alvarez are at the California Institute of Technology, Adams-Cohen is at Stanford University. We thank the John Randolph Haynes and Dora Haynes Foundation for supporting our research. We thank Google for supporting our research through their COVID-19 research grants program.

[1]See R. Michael Alvarez, Nicholas Adams-Cohen, Seo-young Silvia Kim, and Yimeng Li, **Securing American Elections: How Data-Driven Election Monitoring Can Improve Our Democracy**, Cambridge University Press, forthcoming October 2020. Or watch a related presentation online, https://youtu.be/mbEpKcFxp8s.

1. Election day voting: provisional ballot, voting machine, ballot
2. Voter fraud: election fraud, election manipulation, illegal voters, illegal votes, dead voters, noncitizen voting, noncitizen votes, illegal voting, illegal vote, illegal ballot, illegal ballots, dirty voter rolls, vote illegally, voting illegally, voter intimidation, voter suppression, rigged election, vote rigging, voter fraud, voting fraud, vote fraud, vote buying, vote flipping, flipped votes, ballot stuffing, ballot box stuffing, ballot destruction, voting machine tampering, rigged voting machines, voter impersonation, election integrity, election rigging, duplicate voting, duplicate vote, ineligible voting, ineligible vote, ballot harvesting, voter purge, voter purges, voter purging, alien voting, voter coercion
3. Remote voting: absentee ballot, mail ballot, vote by mail, voting by mail, early voting, vote at home, voting at home, safe voting vote safe, ravbm
4. Voter ID: voter identification, voting identification, voter id
5. Polling places: polling place line, precinct line, pollworker, poll worker
6. Election challenge (not currently shown in real-time): blue shift, red mirage, vote count, ballot count, Vote recount, ballot recount, ballot challenge, signature cure, election judge
7. California and Southern California election adminstration (not currently shown in real-time): OCRegistrar, #ocvote2020, #ocvotecenters2020, #protect2020, #OrangeCounty, #OCVotes, LACountyRRCC, lacountyrrcc, #VSAP, #LAVotes, #LACounty, CASOSvote, #VoteCalifornia
8. Hashtags: #VoteAtHome, #VotebyMail, #VoteatHome2020, #Election2020, #DeliverMyVote, #BallotTracking, #VoteSafe, #VotebyMail2020, #2020elections

We will update this methodology note as necessary.