ISSN 2324-7878 (online)
palrap.org

PaLA College & Research Division (CRD)

Pennsylvania Libraries: Research & Practice

Feature

Collecting Pennsylvania Political Twitter Data

Andrew M. Dudash & John E. Russell

Andrew M. Dudash is the Librarian for Political Science, Policy Studies, and Government Information at Penn State University, amd846@psu.edu

John E. Russell is the Digital Humanities Librarian at Penn State University, jer308@psu.edu

During the two most recent elections we have seen the importance of social media, and Twitter in particular, for political discourse. This paper describes the effort of an academic library to collect election-related Twitter data from Pennsylvania-specific organizational accounts and hashtags for 2018 and 2020 in the run-up and aftermath of both election cycles. Because of its importance to understanding contemporary politics and its historic value, libraries need to consider the opportunity to collect and make this data accessible to Pennsylvanians.

Introduction

Pennsylvania has become a battleground state for national elections in recent years that has garnered national attention and increased interest from presidential candidates. In the 2018 mid-term election, 58% of all registered voters participated (DeJesus, 2018) and voter turnout in 2020 broke the previous record (1960) with almost 71% of registered voters participating (Shortell, 2020).

Events of political consequence are being discussed on Twitter and this commentary can provide valuable data for both the public and scholarly researchers. It was particularly evident in the last election cycle that Twitter became a major platform for Pennsylvania political discourse and an important tool for understanding the political moment.

This article describes the effort to collect election-related Twitter data from Pennsylvania-specific organizational accounts and hashtags for 2018 and 2020 in the run-up and aftermath of both election cycles. It is an update to the practice paper "Pennsylvania Perspectives of the 2016 Election," which was a "web and social media archiving effort aimed at documenting the people, voices, moments, and prominent issues in the Commonwealth of Pennsylvania during the final 100 days of the 2016 election" (Pinter et al., 2017). The data-collection effort described here diverges from the original in that no web archiving was conducted and we streamlined the scope of Twitter account collection while also spanning both the 2018 midterm and 2020 general elections.

Over the course of this data collection effort, multiple people at Penn State were involved in the process. Eric Novotny, Jeffrey Knapp, and Andrew Dudash, subject specialists representing history, communication, and political science, curated the Twitter accounts and hashtags before the 2018 collection. Digital Archivist Benjamin Goldman led the 2018 project, and Digital Humanities Librarian John Russell ran the collection efforts in 2020, provided technical support, and supported the creation of image snapshots from the data.

Data Collection

The Twitter data collection criteria was based on the criteria from 2016 but involved more subject- specialist representation and curation. We curated a collection of Pennsylvania-specific organizational accounts (those associated with political parties, advocacy groups, college/university groups, and media organizations with a focus on statewide politics), while eliminating the collection of specific candidates. The tweets collected from the organizational accounts represent the most significant part of the collection, about 80% of the total in both 2018 and 2020 datasets. We selected a handful of Pennsylvania-specific hashtags that were frequently used by our selected Twitter accounts (#padems, #papolitics, #MakeitHapPENN, #VoteLocalPA, #TurnPAblue, #PAJustice, #phillypolitics, #FixPA, #TeamPA, #KeepPARed). To keep the Twitter data collection focused on Pennsylvania, general hashtags (such as #election or #trump) were not used.

The data collection was done using a Penn State-hosted instance of Social Feed Manager (SFM) (George Washington University, 2016). The primary advantage of SFM is you can set up the collection parameters, turn it on, and let it systematically collect tweets over time. However, in both 2018 and 2020, data collection was halted for short periods of time due to minor technical issues on our end as well as restrictions that Twitter places on the number of tweets one can collect per hour. None of these hiccups were major but did require some time to manage and also develop strategies for patching any gaps with additional one-time searches to try to ensure we didn't miss many (or any) tweets.

The timelines for this project were different in 2018 and 2020. In 2018, we started in early summer and let the collecting run a little bit past election day to gather some reactions to the results. In 2020, we wanted to get more of the election run-up and started collecting in March. Because of how the election was perceived by some groups, we let the data collection run until just after the Electoral College vote in December. For each election year, once the tweet-gathering was finished, the datasets were exported as CSV files, then merged into a single file with any duplicates removed. These files were then analyzed to generate visualizations that help summarize the contents. The analysis was done using a lightly modified version of the R program tweetmineR (Shaffer, 2018). This code allowed us to generate a timeline of the tweets gathered, the most frequently mentioned accounts and hashtags, the most frequently mentioned bigrams and trigrams, and the most active accounts in the dataset.

The visualizations of some of the highlights of the Twitter datasets allows one to get a quick overview of the contents of the dataset and permits some very basic points of comparison. In the images below, for example, we can see the most mentioned and retweeted accounts in our datasets and the dominance of Donald Trump in 2020 Twitter discourse.

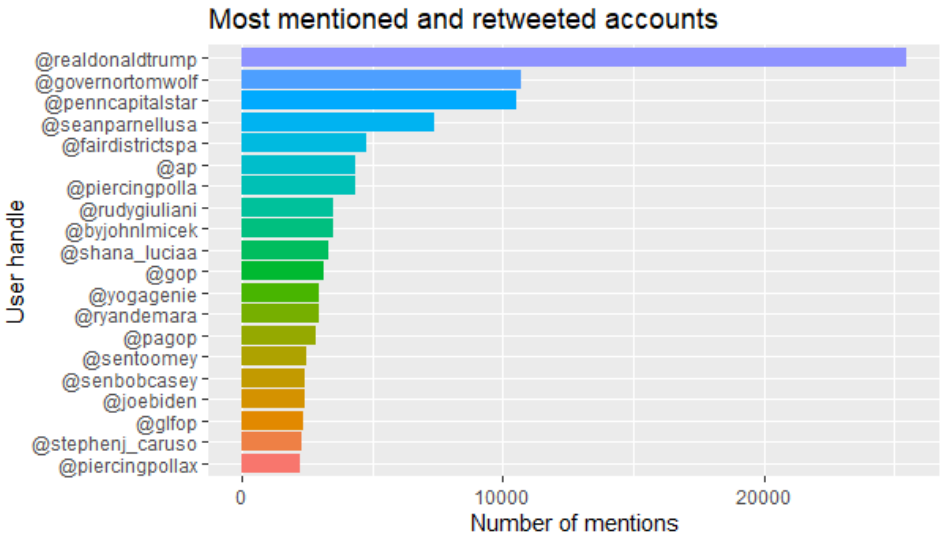

Figure 1

Chart of most mentioned and retweeted accounts from the 2020 Pennsylvania Election Twitter data set.

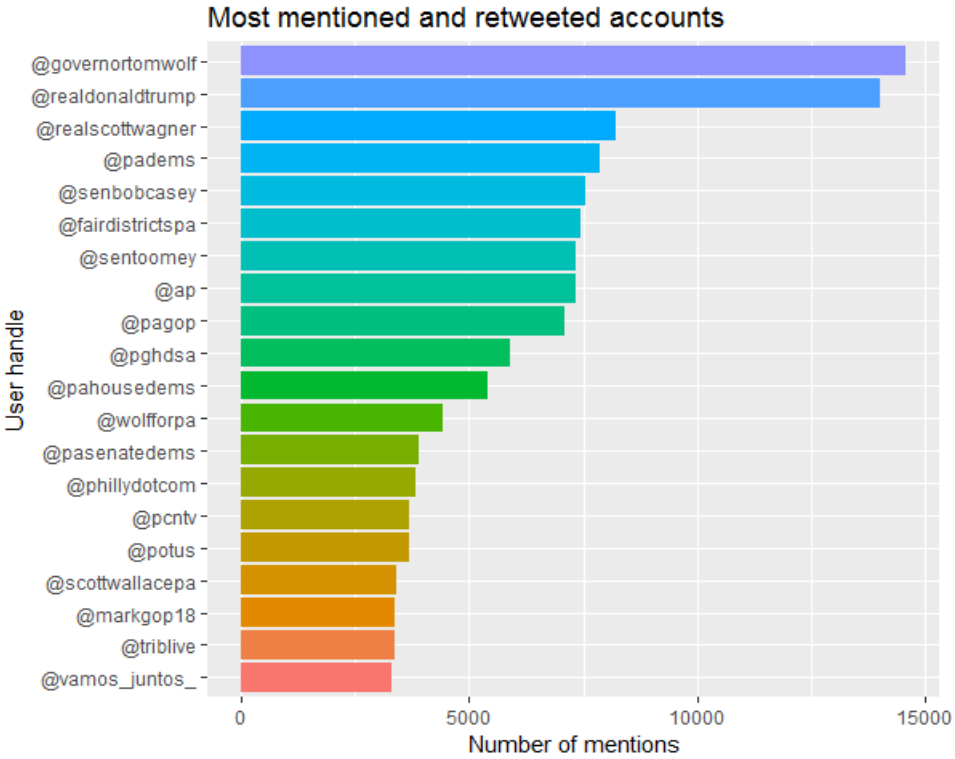

Figure 2

Chart of most mentioned and retweeted accounts from the 2018 Pennsylvania Election Twitter data set.

These are not intended to be for research-level analyses, but they do provide a glimpse into what might be learned from the full datasets and the contents of the dataset.

Access

Our goal with these Twitter datasets is to collect and, to the extent that we are able, provide access to this snapshot of Pennsylvania-related political life. The Twitter rules covering sharing of Twitter data do not allow us to make the full dataset of tweets publicly available but do permit sharing all the unique identifiers (Tweet IDs) of all the tweets that we collected (Twitter, 2021). These IDs can be turned back into the tweets themselves using a program such as Hydrator (Documenting the Now, 2020). For each election year, a plain text file with all the tweet IDs, the summary visualizations, and a README file with information about the data collection are deposited in Penn State's institutional repository, ScholarSphere.

Conclusion

It is important for libraries to consider collecting social media data from Twitter and other platforms. It offers the ability to capture societal events while also making the data available to users. There is also consideration about the collection of public messages from organizations, elected officials, and political operatives, while also looking at how these messages resonate and spread across the information landscape. Future social media data collection efforts specific to Pennsylvania could include the Twitter accounts of Pennsylvania Senate and House officials and how that messaging may shape public opinion or policy. What to collect could also be user driven in some instances and may encourage more engagement across a user base or institution.

Creating collections that capture Pennsylvania perspectives of societal events can help us preserve this unique conversation for historical perspective, but also support research on discourse related to elections and other events that have societal impact. Social media data from the last two election cycles in Pennsylvania should certainly reveal something about these events and the Pennsylvania perspectives that have been shaped in the process.

References

DeJesus, I. (2018, November 19). Midterm voter turnout in Pa. keeps up with historic national levels. www.pennlive.com/news/2018/11/midterm-voter-turnout-in-pa-keeps-up-with-historic-national-levels.html

Documenting the Now. (2020). Hydrator. *GitHub*. github.com/docnow/hydrator

George Washington University Libraries. (2016). Social Feed Manager. *Zenodo*. doi.org/10.5281/zenodo.597278

Pinter, A. T., Goldman, B., & Novotny, E. (2017). Pennsylvania perspectives of the 2016 election: A project to collect web and social media content around significant societal events. *Pennsylvania Libraries: Research & Practice*, *5*(2), 96-106. doi.org/10.5195/PALRAP.2017.146

Shaffer, Kris. (2018). tweetmineR. *GitHub*. github.com/kshaffer/tweetmineR

Shortell, T. (2020, November 17). Biden vs. Trump showdown drove Pennsylvania voter turnout to historic high. *The Morning Call*. www.mcall.com/news/elections/mc-nws-pa-2020-election-set-modern-record-20201117-dook5smfy5emjnzzkvhltoh4vq-story.html

Twitter. (2021). "More About Restricted Uses of the Twitter APIs." *Twitter Developer Terms*. developer.twitter.com/en/developer-terms/more-on-restricted-use-cases