Tracking Changes in Political Discourse on Twitter

Nick McCulloch
American University

Abstract

This paper examines changing word choice in political discussion on twitter and methods for analysis. Identifying linguistic patterns in a chosen sphere and tracking those trends through an evolving context are key challenges for text-miners. This paper analyses political discussion and jargon from 2007-2022 on Twitter. To do so it had to address 1) How to identify "political" tweets and exclude others. 2) How to collect a body of tweets large enough for analysis. 3) How to parse and analyze that data. Analysis of the collected tweets revealed notable attributes in twitter usage, language make-up and issues for future projects.

Keywords: Language, Politics, Twitter.

1. Introduction

Language is constantly in flux. Every year brings with it new words to give passionate expression to the issues of the day. For text analysis, that means trouble. Researchers must determine how best to identify, collect, and analyze a mass amount of text even as, year by year, it continues to grow and change.

In some contexts this is easier than others. Some institutions publish data in structured formats with consistent tagging such as government or research institutions. Others exercise significant editorial control and tightly limit the style for what documents are published, such as newspapers. On Twitter, there's none of that. Any individual can get a twitter account and speak on any topic at any time.

However, despite its relative lack of editorial or topical constraints Twitter offers several advantages for text analysis. First, its sheer enormity is both a challenge and a benefit. There are on average 500,000,000 million tweets per day. A researcher that can manage such a large number of tweets has access to a considerable body of text for analysis. Second, since it emerged as a popular platform Twitter has been the center of significant political interest and discussion. The site is utilized by a broad cross section of individuals, meaning it provides primary text examples of common language and discourse that can't be found elsewhere. Third, although the text of a tweet can be relatively unstructured, the metadata provides a host of useful attributes such as user, date, retweets, and more.

Selecting tweets for study can be difficult. Researchers have developed various methods for limiting their search. These methods can be supervised, meaning a human must provide some level of oversight, input or training data, such as selecting pertinent tweets or keywords, or these methods can be unsupervised, meaning they perform their analysis without human input. These methods offer competing benefits and constraints. Human supervised selection conforms to human judgment and sensibilities, but also creates a bottleneck. Unsupervised topic modeling by contrast is scalable, and can provide unexpected results, for better and worse. However, a researcher decides to select their corpus for analysis, it's only the first step. They must also develop a method for collecting and processing the desired tweets.

Twitter provides an official API for interested researchers which can source and return tweets based on parameters such as 'user-ID,' 'date,' or other quality. Exact parameters must be provided for the API to return a useful result. There are also limits on the rate, size, and number of requests to the API meaning some patience and code-assisted automation is necessary to assemble a sizable collection of tweets.

Once assembled, the corpus must be processed and analyzed to identify trends and provide direction for future study. Generally, analysis is performed with a mixture of mathematical and text -based procedures. There are many available methods. Their process and selection are discussed further below as are the results of the study.

1.1. Project Goal

The goal of this project is to analyze political terminology as it's used over time. This includes new words that rise to prominence, sudden changes in the rates of use for existing terms, and changes in the use of words for particular contexts. These trends provide insights into the state of our political system, the issues that preoccupy political elites, and the language with which issues of national importance are discussed. To do so it identifies a model for selecting political tweets, methods for collection, and tools for analysis, culminating in a time-oriented visualization of word usage and trends.

1.2. Project Problem

There are both technical difficulties in using text analysis over larger periods of time and practical difficulties in studying it. Language is vast so the area of focus was narrowed to political discourse, which is an important subject, and a useful focus for ongoing study. However, determining how to tailor analysis to political speech is a challenge.

Twitter is not particularly popular as far as social media platforms go, but it still averages hundreds of millions of daily tweets and users. Beyond mere size, it hosts discussion on every conceivable issue with many fascinating threads wholly unrelated to the study at hand.

Previous researchers have developed methods for organizing such large quantities of data. These might involve collecting a sample of tweets and sorting them manually to provide training data for an algorithm about keywords and metadata. Another method involves developing dictionaries of key words and phrases which can help differentiate between topics and gauge sentiment. These keywords might be drawn from a selection of example documents, such as papers or books. These methods can be quite useful but present some difficulty when studying a *changing* selection of words and authors. The guidance they provide becomes dated and obsolete. Identifying methods for the continual selection of key words and texts is both a primary goal and challenge for this research.

1.3. Opportunity and Potential Data

Political discourse seems like an ideal candidate because of its ubiquity, cyclicality, and tendency to foster discussion.

To facilitate text analysis, Twitter was selected as the data source. The characteristics of tweets, discrete, limited to 280 characters, and timestamped, should be helpful in the analysis. The first tweet was sent in 2006. Therefore, the maximum potential range of tweets to examine is roughly 16 years. There are about 500,000,000 tweets each day meaning there is a large volume of data to work with. Political figures are frequently active on Twitter, and regular events such as elections ensure continued data over the course of time. The Twitter API provides an invaluable tool for data collection, though there are still practical difficulties in its use.

Finally, after addressing topic modeling and data collection the model will need to perform analysis of the text to highlight what insights can be drawn from this data and process.

1.4. Initial Questions

The initial questions to answer will be how to best target political tweets and commentary and how to find and track keywords, especially new words as they are introduced over time. Next is how to collect that data, process, and store it. Finally, the paper will explore methods for tweet analysis.

1.5. Specific Operationalized Questions

The following is a list of specific operationalized questions:

- Was there variation in the sheer number of tweets during this period?
- What were the most commonly used terms during each month and year under consideration?
- What words or terms gained popularity over the period?
- Does politics, as the area of focus, present itself in the data, either in level of activity or terminology?

1.6. Literature Sources

An initial literature review turned up several papers and sources concerning methods for topic modeling and classification. Some of the early work began well before the explosion in data and computing power seen today. Hans Luhn, who gave a developed Key Word in Context Analysis was born in 1896. Many developments have been made since the work of Luhn and other early pioneers. Some have focused on social media specifically and much has been done to expand the tools available for performing this analysis.

Latent Dirichlet Allocation (LDA) is a method of unsupervised topic modeling. LDA breaks documents into topics or themes and clusters words within them based in part on their connection to other words. It is one method for topic analysis[i].

STM: R Package for Structure Topic Models---A generative topic model (similar to LDA) which provides the added benefit of allowing analysis using document metadata.[ii]

Correlated Topic Models provide a topic modeling approach like LDA. However, CTM takes slightly longer with more difficulty in order to produce better models with more visible relationships as an end product.[iii]

Statistical models have also been applied to the study of neologisms. In their work "Where New Words are Born" researchers examine competing theories of semantic sparsity and frequency growth rates noting that both offer some insight in new word formation.[iv]

"Neologisms in Online British-English versus American-English Dictionaries" the tracking and definition of neologisms by commercial publishers in the US and UK. The paper is less statistical but very oriented towards the study of popular language. [v]

"Political Sentiment Lexicon" is an excellent article by fellow graduate researcher on the analysis of partisan sentiment in text[vi]

"Automated Content Analysis with R: Topic Specific Dictionaries" provides an overview of various topical dictionaries found in the R and the creation and use of new dictionaries with R.[vii]

"Lexicoder Sentiment Dictionary" and the earlier paper "Affective News: The Automated Coding of Sentiment in Political Texts" examine the use of Lexicoder Topic Dictionaries and current usefulness.[viii]

"Sentiment Analysis of Political Communication" investigates the use of crowdsourced keyword dictionaries as a method of supervised topic modeling.[ix]

Pew provided an intro to topic models for text analysis including many of their own efforts in the field and methodology.[x]

The above literature and further independent study highlighted both the potential of algorithmic topic modeling and the difficulties created by inscrutable unsupervised approaches and the labor intensiveness of supervised methods. These weighed heavily on the research approach for this study which faced similar limitations.

2. Data Collection

Data was collected using Twitters official API. Specifically, academic research access was sought and received. That approval was used to access Twitter's APIv2 GET /2/tweets/search/all endpoint which allows collection of tweets from any point in the Twitter archive. This paper drew samples from the year 2007 to 2022. The Twitter Developer Platform provided helpful documentation, tools, and example code which was used in developing our query.

2.1. Twitter Endpoint and Rate Limits

The Twitter GET /2/tweets/search/all endpoint provides access to the full archive of tweets with certain restrictions on the size, frequency, and rapidity of requests. Specifically, it imposes a limit on requests of 100 tweets per request, one query per second, 300 queries per quarter hour, and an upper limit of 10,000,000 tweets per month.

Each request must include parameters for which tweets should be collected. These include unique identifiers such as tweet or user ID and more general characteristics such as date or locale. In addition to the rate limits, the GET request method has a maximum character length of 2048 for search parameters and other necessary code.

For the purposes of this project these limitations meant it was necessary to select a subset of usernames to include in each search and to automate requests so that a large enough corpus could be assembled. Although not in the public domain, these Tweets are generally considered fair use for academic research.

2.2. Political Usernames and Target Dates

Tweets by elected officials were deemed to be inherently political to some degree, especially in the aggregate. As official positions, officials are also present throughout the period in question and geographically diverse.

Finding a centralized list of political accounts proved difficult. Twitter has lists of accounts determined to be government affiliated, the House Press Gallery provides the Twitter accounts for current members of the House of Representatives.[xi][xii] The largest source of data no member names and social media accounts came from a joint database maintained by GovTrack, ProPublica, Maplight, and FiveThirtyEight among other groups and individuals. [xiii] To conform to the 2048-character limit, a subset of users were selected to include in each query. 1,995 usernames were selected in total, representing either the elected representative's official or personal account, their press office, or their campaign. These usernames were ordered at random and placed in groups of 10 to 40 to include in each query.

The target dates were the period between 2007, the year after Twitter was founded, and 2022, the date of this study. When responding to a query, the Twitter API collects tweets in reverse chronological order until it reaches the 100-tweet limit. To ensure a selection of tweets from across the period a random date was generated and supplied with each query.

2.3. Data Structure and Size

The finalized data set consists of 343,978 cleaned and preprocessed tweets from 1105 accounts. These were stored in a CSV file as were the reference lists used for sample selection.

Tweets contain a wide variety of data, which is detailed in the Twitter API v2 data dictionary.[xiv] The relevant fields for this study are '[tweet] id', 'text', 'author_id', and 'created_at'.

- id — a unique identifier for the tweet.
- text — the plain text content of the tweet.
- author_id — a unique identifier for the tweet's creator.
- created at — the date of the tweet.

3. Methodology

Tweets were collected for the data set using the Twitter Developer API v2. RStudio was used for analysis. Although many popular R packages were designed for the previous API these were not fully integrated with the new API at the time of the study. To gather useful quantities of data within the API's restrictions, collection was automated. The general process is described below. The full code is online.

Set-Up and Defining Parameters

- Apply for access to the Twitter API
- Install and load required packages
- Gather list of chosen officials' usernames
- Select desired fields for the request
- Generate a matrix of the desired users/dates

Gather Data

- Define function to make request and save the response.
- Make an initial request

Parse out Useful Data

- Parse the response object (received as JSON)
- Extract and process the tweet, author, and place components of the response, as well as metadata about the response itself
- Merge components into one data frame

Perform quality checks and Save

- Check data quality and consistency particularly column types and order
- Append the data frame to previous results and write to .CSV

Automation

- Create nested for loops to perform previous steps including initial setup, data collection, clean-up, and file creation
- Iterate through all sample dates and usernames
- Test results between queries to protect data integrity and structure
- Watch for bad requests and the rate limit

Analysis

- Perform checks and statistical analysis: including data characteristics, breakdown by user and date, variable quality and summary statistics
- Perform initial text analysis on unfiltered data
- Transform and subset data for further analysis
- Create a corpus, tokens, and document feature matrix including pre-processing and filters
- Extract top words and hashtags
- Extract results by year and other variables
- Track results across dates to identify trends
- Use the results to create stop-words and iteratively improve the analysis.
- Create data visualizations

3.1. R-Packages

Several R packages were utilized to extract, clean, and process the data:

- rmarkdown
- knitr
- httr
- jsonlite
- magrittr
- dplyr
- ggplot2
- tinytex
- readr
- tidyverse
- tidytext
- lubridate
- stringr
- Matrix
- slam
- bench
- rJava
- qdap
- tm
- tidyr
- quanteda
- quanteda.textstats
- quanteda.textplots
- quanteda.textmodels
- ggthemes
- scales

Though several packages were utilized for this research particular functions were distinctly important for interacting with the API, parsing or transforming the results and analyzing the text. Specifically, httr was used to interact with the Twitter Developer API, access HTTP elements of the service, and produce response objects. Jsonlite was used to reformat and parse the JSON response. Dplyr was key to manipulating the data throughout the process. Finally, quanteda provided the bulk of text analysis, transformations, and even certain plots.

4. Limitations

The lack of an exhaustive source for government social media accounts was a limitation. All official accounts for current members were included but members who served since 2007 but left office before the current Congress are not. These former members may have represented a different generation, served in a changing district, or been out of sync with political trends, any of which may have been reflected in their tweets had they been included. There's also no guarantee that every unofficial member campaign or personal account is included which may well shape the formality and tone of the results. An authoritative and exhaustive list for every official or at least every Congress would improve the study.

Within the sample of public officials that are currently serving there is extremely uneven usage. Some of this can be attributed to a transition from campaign to official accounts, which were previously active but now defunct.

Other differences can be attributed to length of tenure. Officials that have been in office for a long time have a significant lead over new members with surprising consequences for the data. For example, Senator Rob Portman of Ohio has been in office since 2010, and is highly active on Twitter. By contrast, Alexandria Ocasio-Cortez assumed office in 2019, nearly a decade later. As a result, Senator Portman is by far the most common result in the dataset. Although changes in sampling were made to test for potential distortions, checks on total number of tweets confirmed the significant differences in twitter usage between officials. The text analysis is broken out by tweet not by user which explains why Senator Portman's native Ohio, made it onto the list of most used hashtags.

Another limitation is the relative recency of social media. Although the company was found in 2006, significant usage didn't begin until around 2008. This is apparent in the data where there are a mere four tweets from 2007. Although compelling for certain uses, Twitter data lacks the historic nature of print media.

Further analysis would have been useful in extracting helpful insights. Namely, although a good assessment was provided of the words, hashtags, and trends of elected officials, these were contrasted with the behavior of regular users, even those in direct communication with officials. Meaning that some of the key words and trends merely reflect society as a whole rather than the peculiarities of elected officials or politics as a sphere. Such analysis would have required the use of another API and other sampling methods which were beyond the scope of this study.

Finally, no compound token or co-occurrence analysis was performed in this study. Compound token analysis would have been helpful in identifying key phrases that may have shed greater light on the particular features of politics than individual tokens. In that same vein a co-occurrence analysis could have demonstrated the relationship between popular hashtags and regular word choice. Some correlations were identified based on year and election cycle, and keyness was examined to contrast wording before and after key dates but further analysis would be beneficial.

5. Findings

Analysis first examined tweet frequency by year. Figure 1 highlights the paucity of information prior to 2008 and significant subsequent growth. This distribution can be explained in part by Twitter's growth in popularity. However, a second explanation is that as former members aren't included even those active on twitter and so the data set is biased towards the current elected members of Congress.

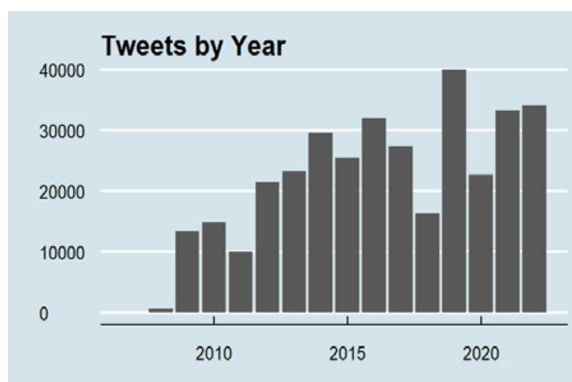

Figure 1. Number of Tweets by Year

5.1. Misleading November Usage Spike

The next finding of interest can be seen in Figure 2, which graphs twitter usage by month summed across all years in the study window. As can be seen below there is a notable spike in usage during the month of November and a slump through the rest of the Fall.

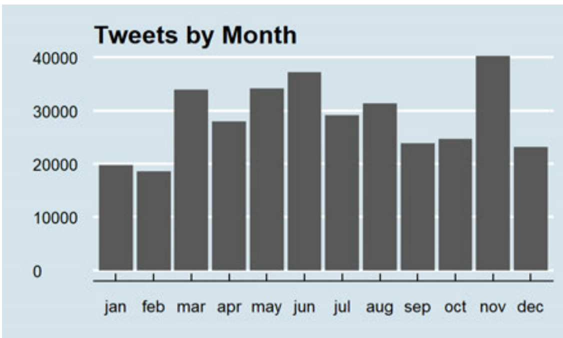

Figure 2. Number of Tweets by Month

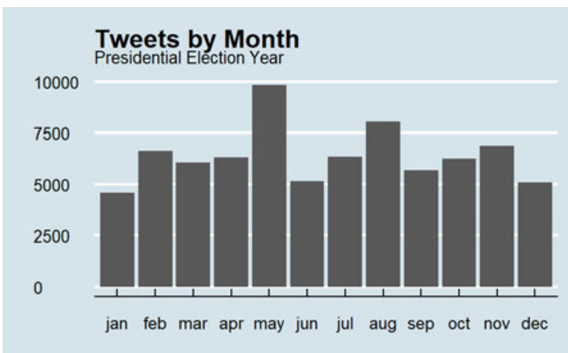

Figure 3. Monthly Tweets---Pres. Election Years

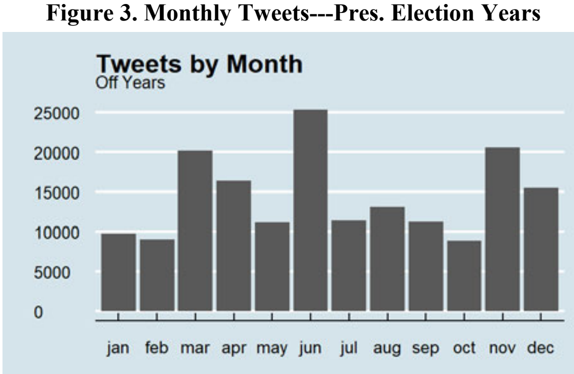

Figure 4. Monthly Tweets---Of Years

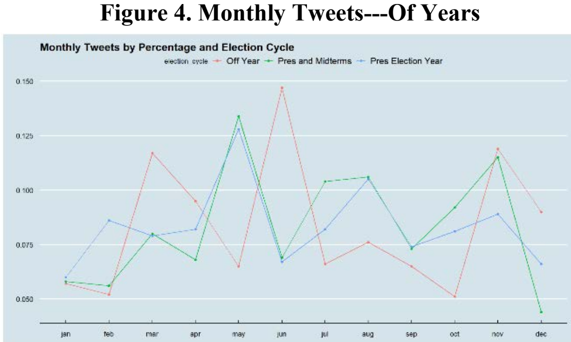

Figure 5. Monthly Tweets---All Election Cycles

This being a political data set a reasonable hypothesis could be made that a November spike represents energetic campaign activity by politicians and their staff in advance of the election. However, when broken out by election cycle the data reveals the most disproportionate November usage occurred during off-years between elections, and that Fall usage was generally more consistent during election years.

5.2. Confirming Unequal Usage by User

The next finding was indicated in the summary statistics and confirmed subsequently. The distribution of tweets by user was unequal. At the high end was Senator Portman with 1056 tweets, and at the bottom were accounts with zero tweets (generally because another account was preferred) Besides the maximum and minimum values of 1056 and 0, the average number of tweets was 19.46, with 1,995 unique user ID's, and 343,978 unique tweets.

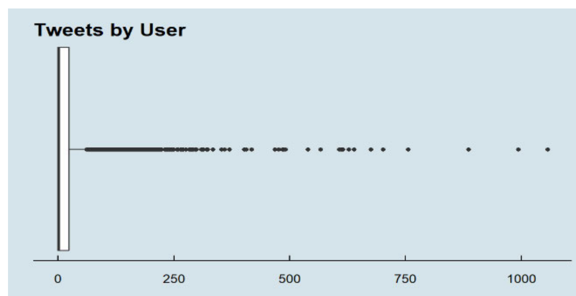

Figure 6. Number of Tweets by User

5.3 Preliminary Analysis and Stop Words

Preliminary text analysis confirmed the importance of stop word dictionaries and processes for filtering out grammatically useful but uninformative language. The most commonly used words before applying any filter were, "the, to, and, of, in, for, a, on, is" and in a nod to Twitter, "rt" for "retweet."

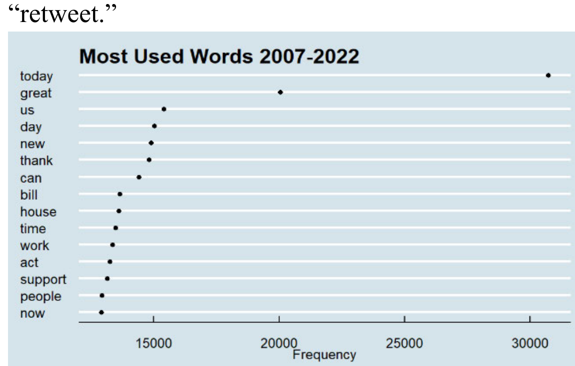

Figure 7. Most Used Words, 2007--2022

5.4. Top Words Hint at Political Focus

Although the words are not contrasted with more general conversation on Twitter the top words for the study timeframe give an impression of their government-oriented nature. As can be seen below, “bill,” “house,” and, “act,” made the top 10 most common terms.

The remaining terms, although not explicitly political, do seem unusual outside of a political context. “Support” for instance is a word frequently used in relation to causes movements and doesn’t seem to warrant such high usage in normal conversation. Confirming these suspicions is an area where more comparative analysis would be useful.

5.5. Most Common Words by Year

When top features are broken out by year there are interesting results as well. The only period where a hashtag makes it to the top features list was #tcot in 2008-2011. That time period is also the only time that a technology-related term made the list with “video.” It was included on of only two issue-based words with “health.” In the later periods from 2016-2019 and 2020-2022, “work,” can be found, although

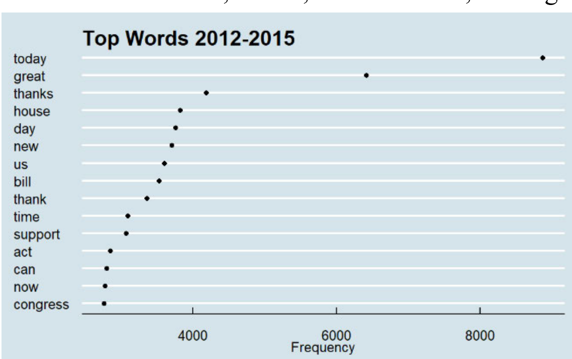

Figure 8. Most Used Words, 2012–2015

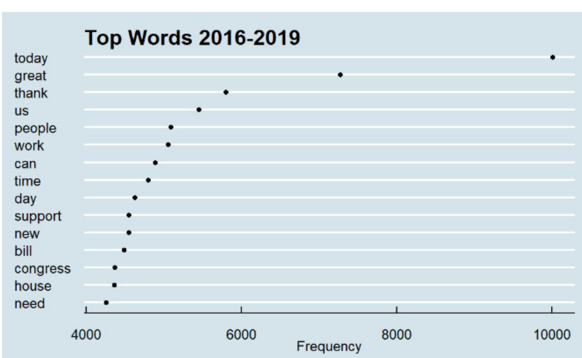

Figure 9. Most Used Words, 2016–2019

whether this is in reference to unemployment or a general call for effort is unclear without further contextual analysis. The persistent popularity of, “today,” across periods is notable and perhaps indicative of Twitters role in breaking news.

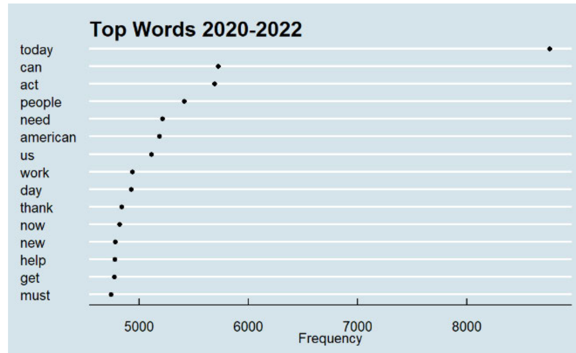

Figure 10. Top Words, 2020--2022

Apparent in the data though unexplained is the significant spike in usage around 2019 which can be seen across all words in the graph below.

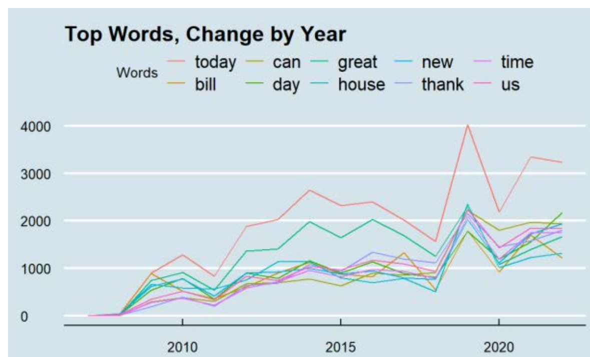

Figure 11. Top Words, 2007--2022

5.6. Hashtags Reveal Topic

A notable finding, that may be peculiar to Twitter data can be observed by hashtag analysis. By design hashtags are used to label and sort information by topic or theme. Unlike the body text of tweets, it is rare for a hashtag to be found on a stop list and many are both topically specific and unique to a particular moment.

As can be seen in figure 12, the theme and context of these tweets is far more readily apparent than the similar graphs of top words to the left. For example, the top five words, today, great, thanks, house, and, day, are far less informative than, the hashtags, #TCOT, #Obamacare, #COVID19, #GOP, and #Jobs. Hashtag use might also reduce comparative use of topical words in the body of tweets compared to mediums where tags aren't found.

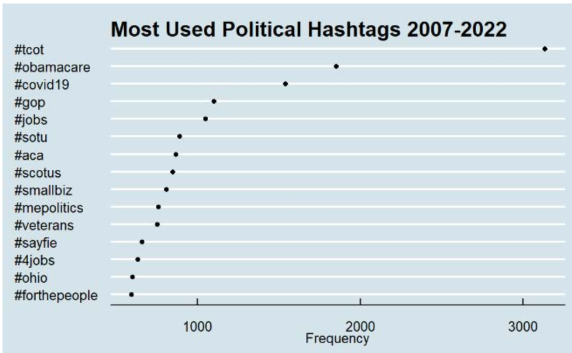

Figure 12. Most Used Hashtags, 2007–2022

5.7. Hashtags Illustrate Time and Place

Another facet of hashtags is how readily understood they are as being tied to a particular current event, place or issue. For instance, the spike in #COVID19 in 2020. Or to provide another example, #HCR, #Obamacare, #ACA, and #healthcare, were all trending around the passage of the Affordable Care Act.

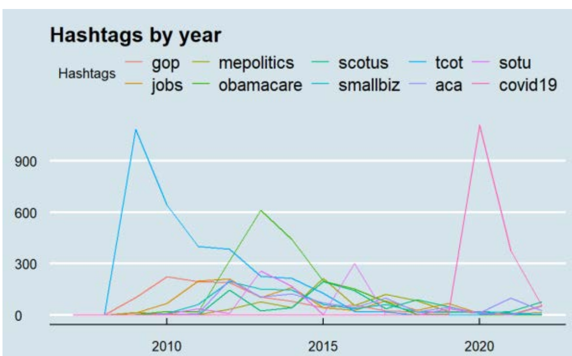

Figure 13. Most Used Hashtags, 2007–2022

5.8. Hashtags Used to Denigrate or Support

Hashtags are also notably used to draw attention to legislative activity by connecting a proposal to a political opponent, perhaps with an unflattering name, #Obamacare, #Trumpcare, #GOPTaxScam, #Bidenflation. Conversely, hashtags might be used to garner support through explicit legislative references, #AmericanRescuePlan, #ChildTaxCredit, #CutCapBalance.

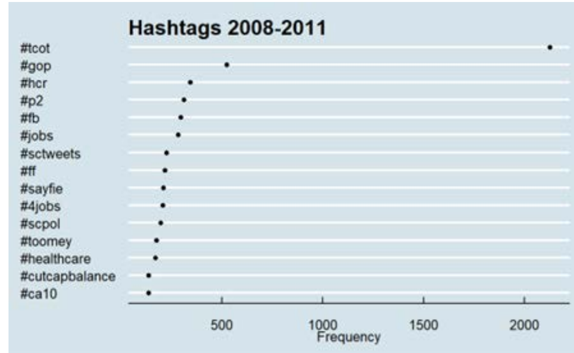

Figure 14. Most Used Hashtags, 2007–2022

5.9. Hashtags as Geographic Labels

Geographic metadata was fairly uncommon in the dataset. This was true to the degree that try-except, code was written to manage its frequent absence. By contrast hashtags are frequently used to indicate the tweets' relevance to a particular locale (often an electoral district). For example, hashtags included, #CA10, #VA10, #Ohio, #AR3, and #TXSen.

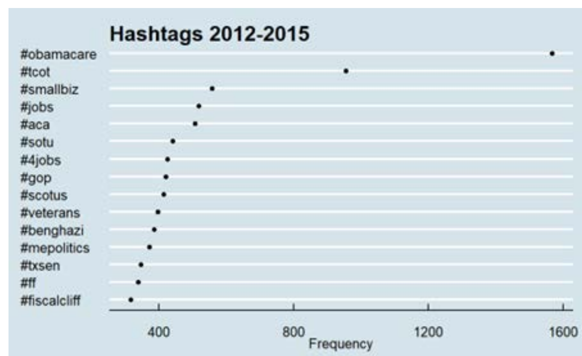

Figure 15. Most Used Hashtags, 2012–2015

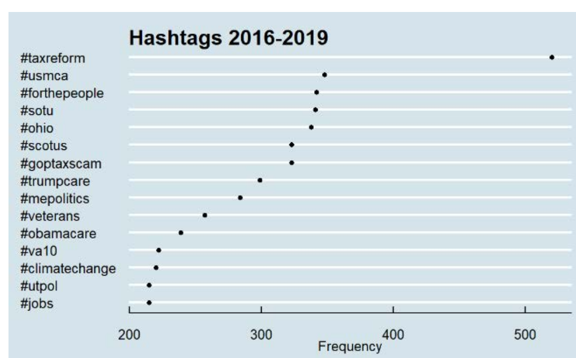

Figure 16. Most Used Hashtags, 2016–2019

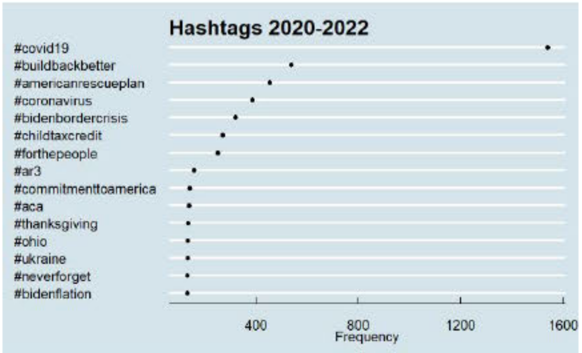

Figure 17. Most Used Hashtags, 2007–2022

5.10. Keyness Offers Clearest Benefits

Although the charts above offer their share of insights, they are outshined by keyness analysis and plots. Keyness analysis, by contrasting features based on both overall frequency and frequency differences when grouped by other variables offers a useful tool for drawing out terminology that is particularly distinctive.

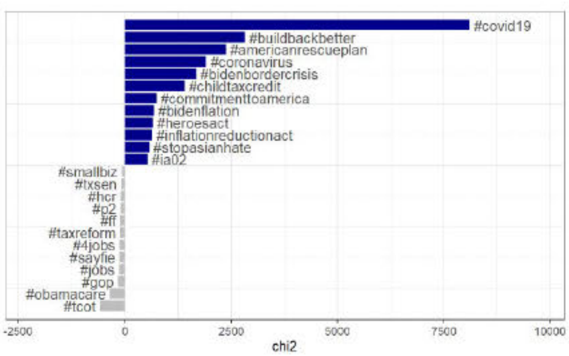

Figure 18. Keyness Plot—Top Hashtags Pre/Post COVID

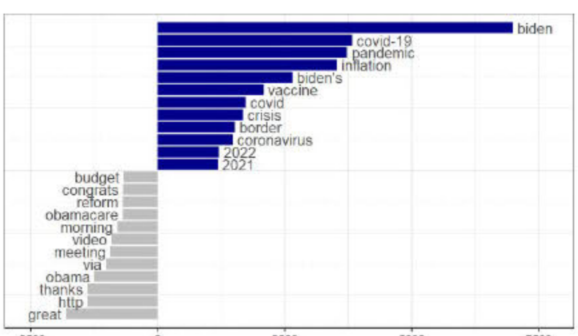

Figure 19. Keyness Plot—Top Words Pre/Post COVID

6. Conclusions and Future Research

The project demonstrated the potential for significant data collection using the Twitter API. It also highlighted the benefits of pairing structured social media, and a predictable assortment of users with topic modeling methods. The analysis of vast Twitter data can be greatly assisted by unsupervised and semi-supervised topic modeling and classification methods. The scale of the Twitter corpus, its topical breadth, and the rate of new tweets make tools which require minimal labor from researchers essential.

At the same time, the literature review highlighted how unsupervised models can benefit from the ample metadata provided with Tweets, as in the STM method. It highlighted the potential for crowd-sourced tagging as a method of supervised modeling and classification. The benefits of large data streams with in combination with voluntary tagging could prove substantial, particularly with an eye towards the long term where data will accrue at a rate faster than researchers are likely able to track and parse with out some form of assistance.

Twitter as a platform also presented peculiar characteristics which seem beneficial but may not lend themselves to other spheres. For instance, hashtags find some use in certain social media platforms but they are not universal. Furthermore, although politicians tend to be active on Twitter, they aren't found on every other platform to the same degree.

The relatively bland results from simple keyword analysis reinforced the necessity of more advanced models such as keyness, co-occurrence and compound token analysis. Without these the benefits of this data set seem more limited.

Adding to that concern, the lack of official social media directories is an unhelpful impediment to public research. Nor is Twitter the only medium where it is likely to prove challenging, other platforms likely face many of the same issues especially where they aren't tied as directly to a single, verifiable public persona.

Further research should also further investigate the use of time-based analysis for known fields as many of the more interesting results were revealed by changes over time rather than absolute numbers at any one moment.

7. References

[i] Blei DM, Ng A, Jordan M (2003). “Latent Dirichlet Allocation.” Journal of Machine Learning Research, 3, 993–1022.

[ii] M.E. Roberts, B.M. Stewart, D. Tingley, “STM: R Package for Structural Topic Models,” Journal of Statistical Software, doi:10.18637/jss.v000.i00

[iii] D.M. Blei, J.D. Lafferty, “A correlated topic model of Science.” The Annals of Applied Statistics, 1(1), 17–35. (2007)

[iv] M. Ryskina, E. Rabinovich,; T. Berg-Kirkpatrick, D.R. Mortensen,; and Y. Tsvetkov, "Where New Words Are Born: Distributional Semantic Analysis of Neologisms and Their Semantic Neighborhoods," Proceedings of the Society for Computation in Linguistics: Vol. 3 , Article 6. (2020)

[v] S. Creese, “Neologisms in Online British-English versus American-English Dictionaries” 18th Euralex International Congress, Lexicography in Global Contexts (2018)

[vi] C. Vail, “Political Sentiment Lexicon,” Rpubs, (2017)

[vii] C. Puschmann, M. Haim, “Automated Content Analysis with R, Topic-specific dictionaries.”

[viii] A. Quinn, J. Sevenans, S. Soroka, “Lexicoder Topic Dictionaries” Political Texts, Political Communication, 29:2, 205-231. (2012). “Lexicoder Sentiment Dictionary” online at Quanteda.io (2015)

[ix] . M. Haselmayer, J. Marcelo, Sentiment analysis of political communication: combining a dictionary approach with crowdcoding” Qual Quant 51, 2623–2646 (2017)

[x] P. Kessel, “An intro to topic models for text analysis” Medium (2018)

[xi] “Members’ Official Twitter Handles” U.S. House of Representatives, Press Gallery, online at: https://pressgallery.house.gov/member-data/members-official-twitter-handles. (2022)

[xii] @TwitterGov, “U.S. Senate: personal and govt verified accounts.” Twitter, (2022)

[xiii] “Congress-Legislators” Joint Project by GovTrack, ProPublica, Maplight, FiveThirtyEight, online at: https://github.com/unitedstates/congress-legislators. Accessed (2022)

[xiv] Twitter API v2 data dictionary, accessed Oct. 3, 2022, online at: https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/tweet