W200 - Project 2
Proliferation of Fake News

Github Repo: https://github.com/UC-Berkeley-I-School/mids-200-project-2-Kenz-JM
Data:
https://drive.google.com/drive/folders/1fSD69QJe_6D0EL3nH74EXhaGBca8PpId?usp=sharing

JM Stilb, MaKenzie Muller

Introduction:

Information is useful and dangerous. With the proliferation of mobile devices and social media, people often seek out their news online. Oftentimes we share and discuss this information with others, who share it with others, so on and so forth. The speed with which news can be spread and misconstrued contributes to the complexities of today's Internet information ecosystem, and are not well understood despite their severe consequences. We set out to explore these relationships in order to gain a better understanding of distribution dynamics, feedback loops, and how influential figures can amplify certain information. Specifically, we sought to analyze the symbiotic relationship between conservative news outlets and how an influential figure like President Trump amplifies certain news topics.

Research Questions:

We focused on the following research questions in our analysis:

- What sort of symbiotic relationships exist regarding the propagation of information between news outlets and Trump?
- Where does dangerous information originate?
- How quickly can dangerous information be amplified on the Internet?
- Can influential figures dictate what is spoken about on politically-unaligned networks and social media?

Datasets and Data Structure:

Primary Datasets:

Our primary datasets consisted of two groups of csv files. The first was a set of the former President's tweets. Trump's tweets from May 2009 through mid June 2020 were scraped from Twitter and consolidated into a spreadsheet of ~43,000 rows. Each row contains the contents of the tweet as it was written, as well as columns delineating it's URL, number of

Retweets, number of Favorites, mentions of other Twitter users, hashtags, and the date the Tweet was posted.

File: *realDonaldTrump.csv*
Source: https://www.kaggle.com/austinreese/trump-tweets

Data Structure - Trump Tweets:

- Tweet ID (a unique identifier for the tweet)
- Link (link address to the tweet)
- Content (the text written in the tweet)
- Date (the date the tweet was sent, in a MM/DD/YYYY HH:MM format)
- Retweets (the number of times the tweet was retweeted)
- Favorites (the number of times the tweet was 'favorited' or 'liked')
- Mentions (the twitter username mentioned in the tweet, denoted with an @Username format)
- Hashtag (any hashtags used, denoted with #Hashtag format)

Our second primary dataset is a collection of news articles across stations and publishers. News sources include the New York Times, Breitbart, CNN, Business Insider, the Atlantic, Fox News, Talking Points Memo, Buzzfeed News, National Review, New York Post, the Guardian, NPR, Reuters, Vox, and the Washington Post. The aggregator of this data specifically tried to collect from a variety of political leanings as well as a mix of print and digital publications. This dataset consists of ~145,000 records split across three csv files, with each row representing a single article.

Files : *articles1.csv, articles2.csv, articles3.csv*
Source: https://www.kaggle.com/snapcrack/all-the-news?select=articles1.csv

Data Structure - News Articles:

- ID
- Database id (unique identifier for the database the data was pulled from)
- Article title
- Publication name (the name of the media network)
- Author name
- Date of publication (the date the news article was published)
- Year of publication
- Month of publication
- Url (link to the news article)
- Article content (the copy from the news article)

Supplemental Datasets:

We supplemented our primary datasets with information from Google Trends and Twitter Trends. Google Trends allowed us to search and export information surrounding the Google search frequency of a specific term or phrase into a csv file. Twitter Trends Storywrangling site lends itself to exporting similar trends in JSON format. In order to determine if one of Trump's tweets prompted an uptick in search trends or news articles, we planned to compare key fake news words with their popularity on Google and Twitter using these sources.

Google Trends - https://trends.google.com/trends/?geo=US

- CSV format
- Joined on Date columns

Twitter Trends - https://storywrangling.org/

- JSON format
- Join on Date columns

Initial Data Exploration and Cleansing:

We began our research by formulating a list of keywords infamously associated with Trump and his Twitter account. While our code would work with any keyword, we focused on the following words and phrases:

Keywords for Analysis

Crooked Hillary Clinton / Crooked Clinton / Crooked Hillary
China Virus / Chinese Virus / Kung Flu
Fake news
No collusion
Witch hunt
Lamestream media

Data Cleansing and Challenges:

As mentioned previously, we planned to compare the proliferation of these keywords among Twitter, Google, and news articles as a result of Trump's social media behavior. From here, we developed several functions to transform our datasets into dataframes from which to extract meaningful pieces of information. Several data wrangling and cleansing techniques are described below, with full function code in the accompanying jupyter notebooks.

- We formatted Google trend csv files and Twitter trend JSON files into dataframes to better visualize search popularity and frequency, respectively.
- We converted all dataframe dates in the YYYY-MM-DD format for ease of use and consistent filtering across datasets.
- We developed a keyword search and count function to locate a given keyword inside of a specific column of a dataframe.
- We created a function to group a dataframe based on weeks to plot Twitter trends in a more straightforward manner, as they were originally ranked by days.
- We wrote a function to return a subset of a dataframe based on a specified start and end date based on the timing and usage of a given keyword.
- We accounted for various versions of keywords by converting all strings to uppercase. With more development, we could also have accounted for misspelled words with the use of wildcards.

Other Considerations:

- Twitter trends use ‘rank’ to describe n-gram popularity. N-grams are phrases made up of 1, 2, or 3 word phrases within the Storywrangling site, and are ranked each day based on usage on the website. A lower rank means the word is more popular, whereas rarer words have a higher rank. For example, ‘Black Lives Matter’ was ranked as the 4th most popular phrase on Twitter on Jun. 2nd 2020, whereas ‘hahaha’ hovers consistently around the 4,000 - 3,000 rank. Ranks go up to 1 million before the n-gram is cut from the Twitter trends data collection.
- We opted to include retweets in our Twitter data sources in order to show the social amplification that can occur on social media.
- The article.csv data contains a column with the full news article for each record. This column surpassed the standard character limit for a csv file, and had to be manually extended in order to continue working with the dataset.
- COVID-19 is recent as of 2020, and therefore was not included in the news articles datasets. Therefore, we focused our virus fake new term on Google and Twitter trends.

Graphical Analysis and Figures:

We categorized our analysis into two sections. The first consists of keyword exploration between Trump’s tweets and Google and Twitter trends from the same time period. We broke down each term into its own set of plots - one to compare to each type of trend. The main *realDonaldTrump.csv* records were plotted against records from the supporting datasets.

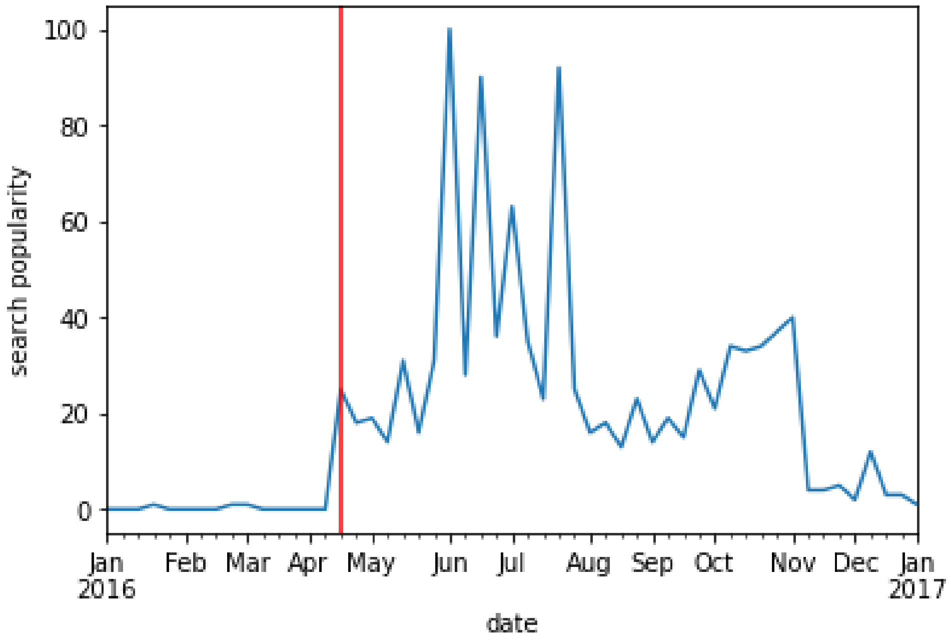
Using the Google trends user interface, we created and downloaded a csv file of the Google search data for each term over a 5 year period. For Twitter trends, we used the Storywrangling interface to create term-specific JSON files. After turning these supplementary sources into dataframes, our date range function was applied to shorten the date window onto a specific time period for plotting purposes. Graphs of Google and Twitter trends were overlaid with vertical lines indicating a Trump tweet, in order to showcase the time series relationship between the sources for a given keyword or phrase.

Our second category of analysis dives deeper into how news articles contribute to the spread of fake news, again defined as the frequency of the items on our list of keyword terms. We transformed the main dataset of news articles into data frames from which we plotted overall use of a specific phrase and then their alignment with Trump's tweets containing the phrase around the same time period.

Keyword Analysis:

A. 'Crooked Hillary' Term Analysis

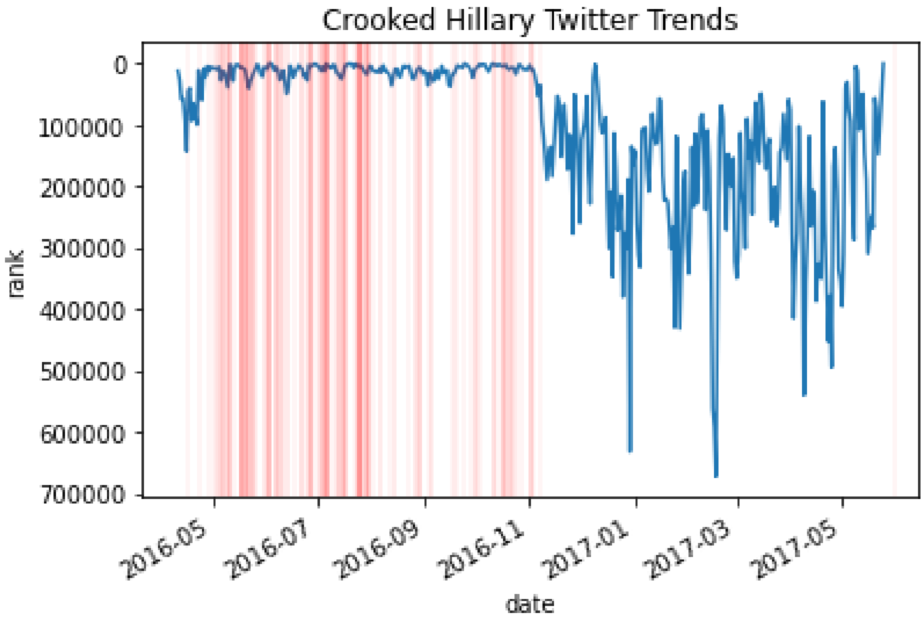
Crooked Hillary Google Searches



Description and Findings:

The figure above shows the Google trend for 'Crooked Hillary' searches within the US from Jan. 1st, 2016 to Jan. 1st, 2017 in blue. Trump's first tweet with the phrase 'Crooked Hillary' is plotted as the vertical red line.

Donald Trump demonstrates a strong ability to create and amplify trends for media bytes such as "Crooked Hillary". He first tweeted "Crooked Hillary" in April 2016 and since in the following months, search popularity skyrocketed over the next several months before dropping back down after the November election.

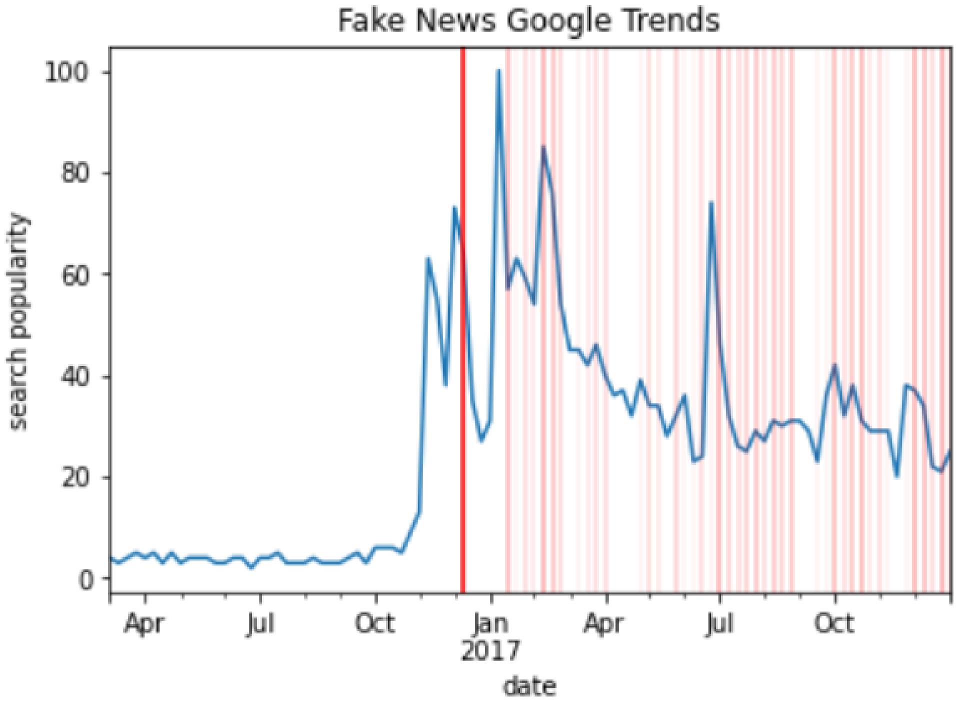


Description and Findings:

The figure above shows the Twitter ranking for the use of ‘Crooked Hillary’ within the US from Jan. 1st, 2016 to Jan. 1st, 2017 in blue. Trump’s tweets using the phrase ‘Crooked Hillary’ are plotted as the vertical red lines.

Donald Trump also has the ability to sustain trends. As discussed, the red lines signify all of Donald Trump's tweets containing "Crooked Hillary" during this time period. These tweets align with the highest reported months for "Crooked Hillary" trends on Twitter. This doesn't necessarily imply causation, but there is a high correlation between the two indicating that he is at least using his influence to contribute to the phenomenon.

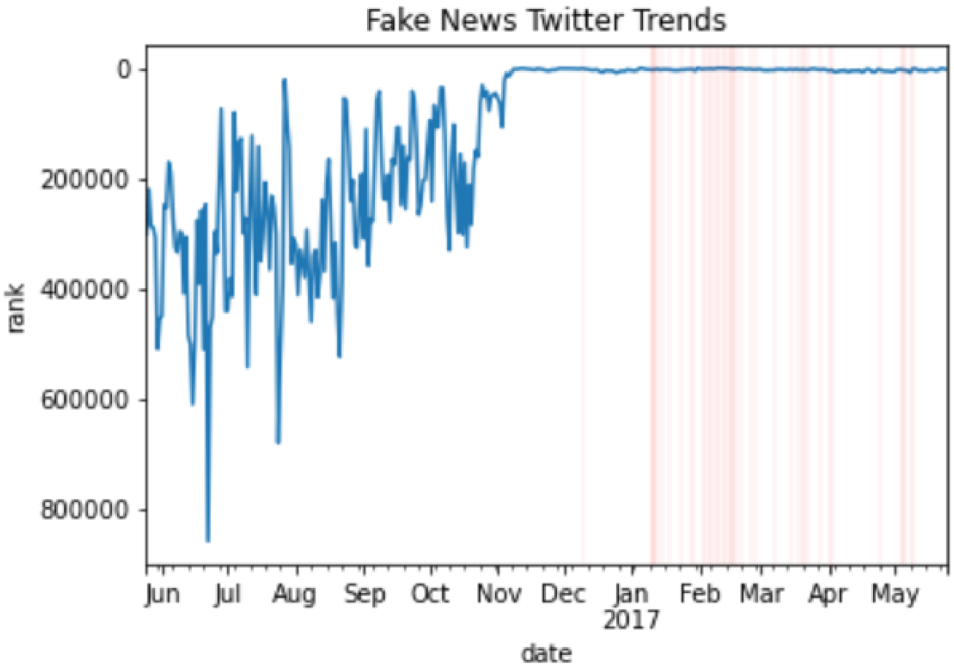
B. ‘Fake News’ Term Analysis



Description and Findings:

The figure above shows the Google trend for ‘fake news’ searches within the US throughout 2017 graphed as the continuous blue trend line. Trump’s first tweet with the phrase ‘fake news’ is plotted as the bold vertical red line with his subsequent ‘fake news’ tweets shown as the fainter red lines. The bolder the line, the more Trump tweeted about the given phrase.

"Fake News" is an interesting case due to the sheer amount of precedence for the term prior to Donald Trump’s first tweet about it. However, there is a strong case to be made that Donald Trump popularized the phrase. His initial tweet in early December of 2017 is an immediate precursor to a clear spike in Google searches involving the term. As you notice, there is a slight spike prior to Donald Trump's first tweet, but it is important to note that Donald Trump has other media outlets, especially as this was the time of the 2016 election, to utilize. It is also clear that many of Donald Trump's tweets in the following months correlate with spikes in searches. It is worth mentioning that these are the first few months of his presidency.

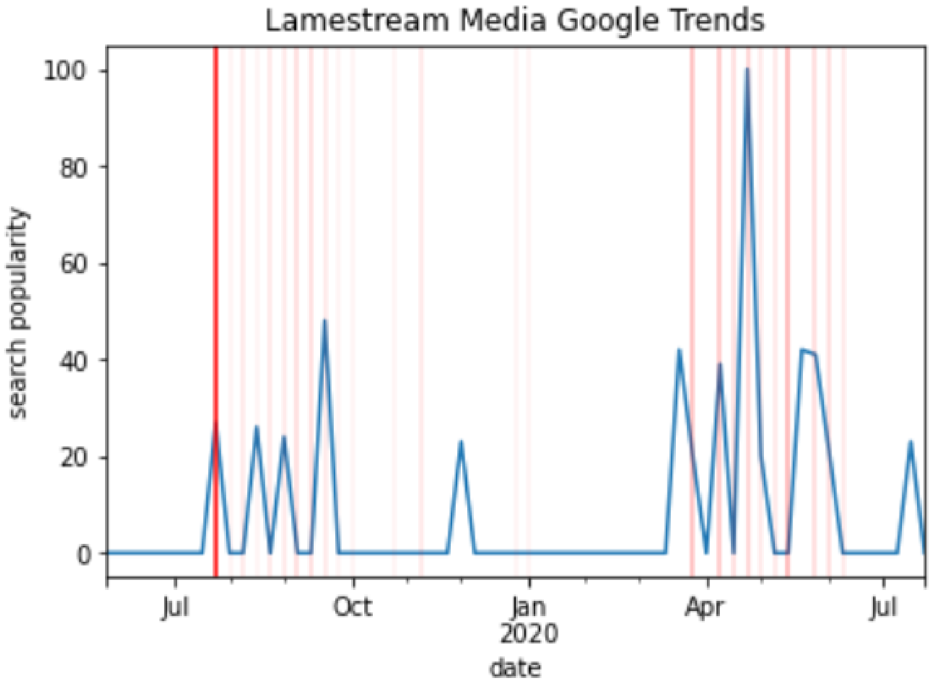


Description and Findings:

The figure above shows the Twitter frequency rankings for the use of ‘fake news’ within the US from Jan. 1st, 2016 to Jan. 1st, 2017 in blue. Trump’s tweets using the phrase ‘fake news’ are plotted as the vertical red lines.

The Twitter trends for "fake news" began their spike in ranking in November 2016, the month of the Presidential election, and maintained their top-ranking throughout the following months. Again there is no clear indicator that Trump caused this sustained spike in rankings, but there is strong evidence that he contributed to its perpetuation of this phrase and its pervasiveness on Twitter.

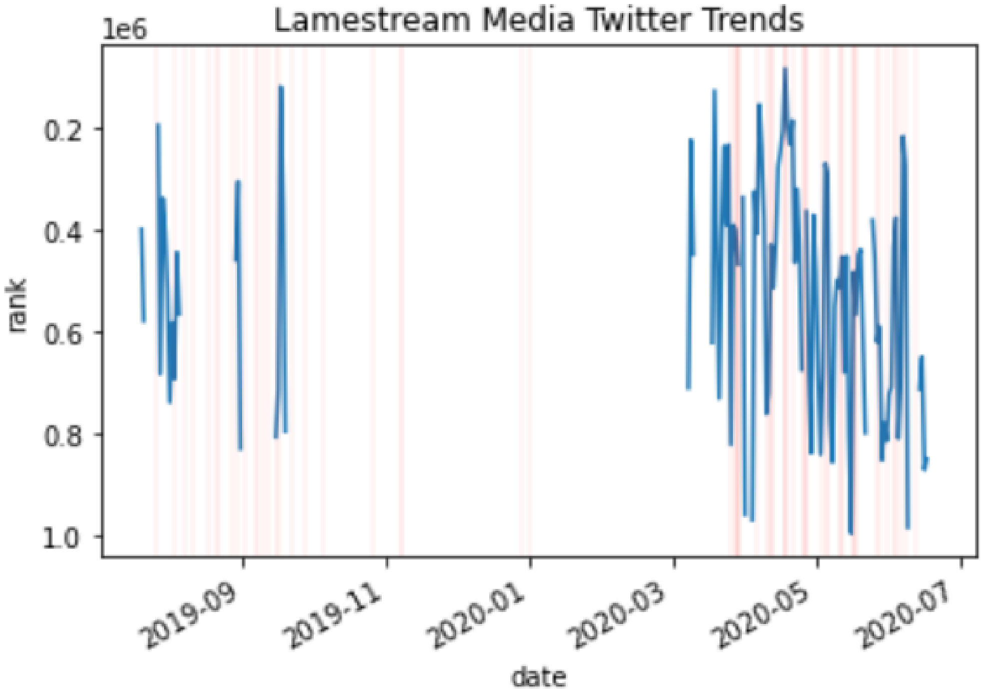
C. ‘Lamestream Media’ Term Analysis



Description and Findings:

The figure above shows the Google trend for ‘lamestream media’ searches within the US throughout 2020 graphed as the continuous blue trend line. Trump’s first tweet with the phrase ‘lamestream media’ is plotted as the bold vertical red line with his subsequent ‘lamestream media’ tweets shown as the fainter red lines.

"Lamestream Media" appears to be a media byte that didn't resonate as some of the others Trump has used during his time. His first tweet does appear to be the catalyst for the spikes in Google searches and he tweeted it during many of the same months of the spikes in Google searches. However, this media byte never really maintained the same level of magnitude as others have.

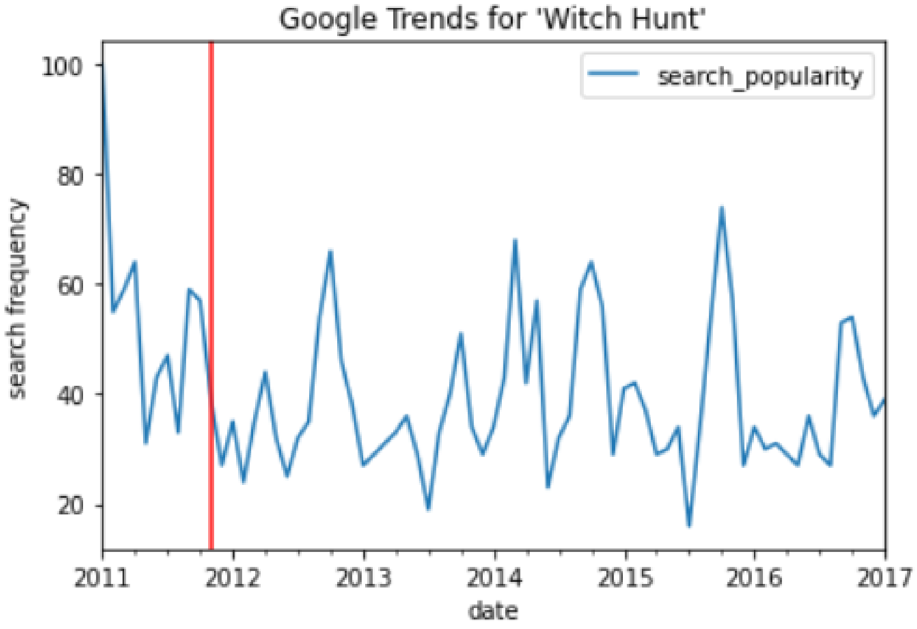


Description and Findings:

The figure above shows the Twitter frequency rankings for the use of ‘lamestream media’ within the US from Jun. 1st, 2019 to Jun. 30th, 2020 in blue. Trump’s tweets using the phrase ‘lamestream media’ are plotted as the vertical red lines.

Again, "Lamestream Media" appears to be a failed attempt to create a viral media byte. Trump had tweeted the term around the same time as the most prevalent Twitter trends, but these trends were weak and short-lived (plus the data is rather noisy). This demonstrates that while Trump's influence may be powerful, not every attempt to guide the national discussion is successful.

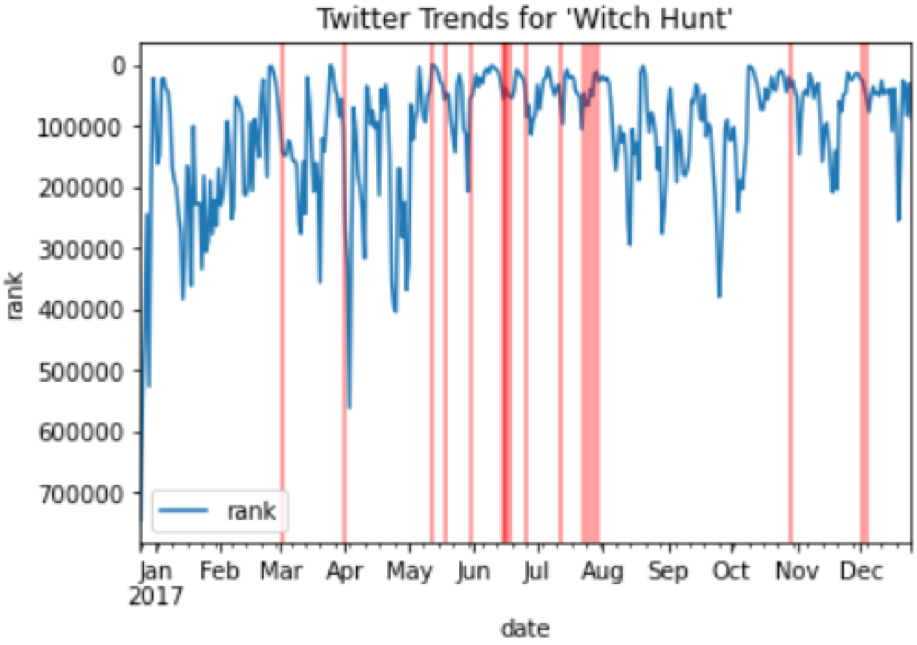
D. ‘Witch Hunt’ Term Analysis



Description and Findings:

The figure above shows the Google trend for ‘witch hunt’ searches within the US from 2011 to 2020 graphed as the continuous blue trend line. Trump’s first tweet with the phrase ‘witch hunt’ is plotted as the bold vertical red line.

‘Witch hunt’ had a much earlier debut than some of the other key terms, but it served the same purpose. Trump first tweeted this phrase on Nov. 4th, 2011 to share an interview he did about Herman Cain. The tweet read, “My interview on @gretawire discussing the economy and @TheHermanCain "Witch Hunt" http://bit.ly/umKRY4.” While long before his time as a prominent political figure, we see that this tweet falls on the graph directly after a spike in the Google trends like for ‘witch hunt.’ This may suggest that Trump saw the term being used elsewhere and adapted it into his own social media presence.

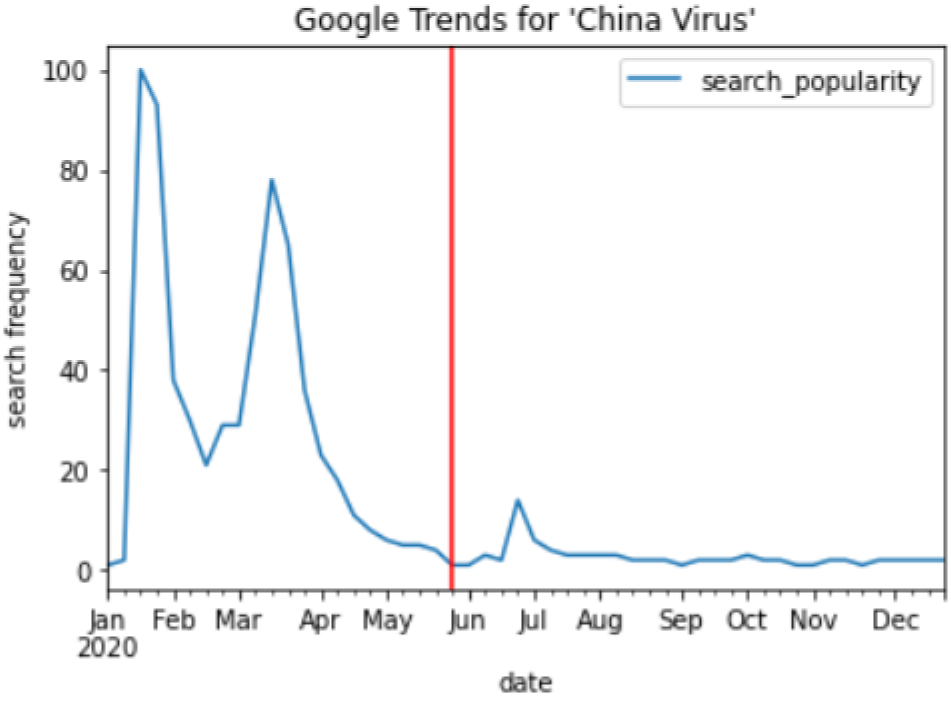


Description and Findings:

The figure above shows the Twitter rankings for ‘witch hunt’ searches within the US throughout 2017 graphed as the continuous blue trend line. Trump’s tweets with the phrase ‘witch hunt’ are plotted as the vertical red lines.

It is evident that the phrase ‘witch hunt’ remained relatively prominent on the Twitter rankings despite fluctuations. Most notably, from May to August, these words stayed highly ranked, coinciding with Trump’s most frequent tweets on the subject. This correlation suggests that there is indeed a connection between Trump’s behavior and that of the site as a whole in regards to this phrase.

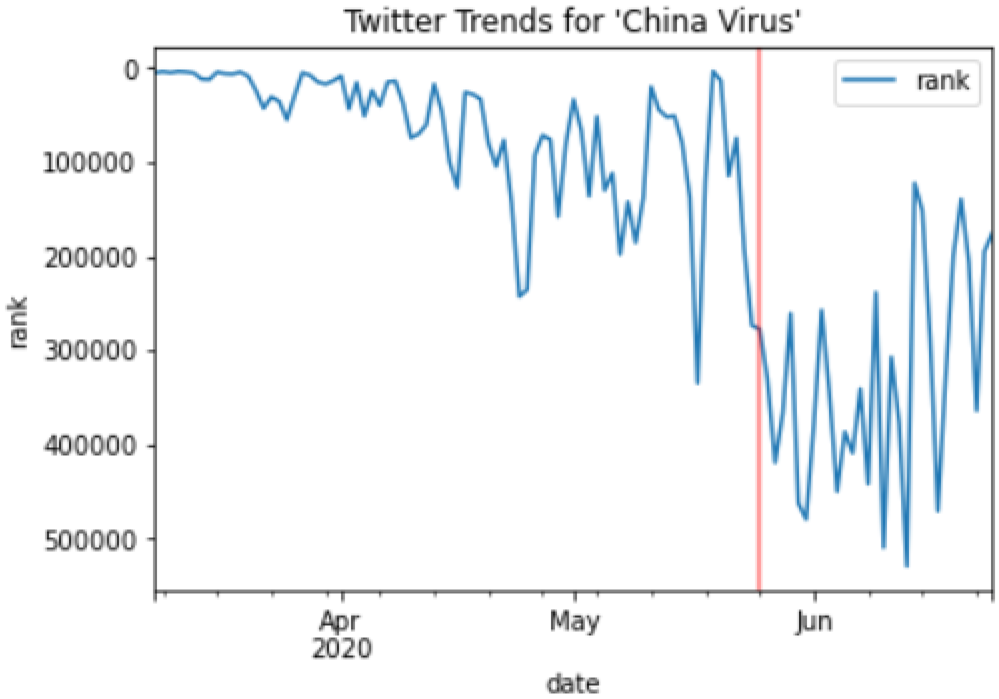
E. ‘China Virus’ Term Analysis



Description and Findings:

The figure above shows the Google trend line for ‘china virus’ searches within the US throughout 2020 graphed as the continuous blue trend line. Trump’s first tweet with the phrase ‘china virus’ is plotted as the vertical red line.

There are notable large spikes in Google search trends for this keyword likely due to the vagueness of the phrase and unknown atmosphere of COVID-19 at the beginning of 2020. Similarly, it is not uncommon for average users to type two keywords that are only tangentially related to one another in order to bring up vast Internet search results. It’s possible that Trump’s first tweet referring to the name of the virus as the “China Virus” is only loosely related to overall Google trends for the given year based on our dataset, as their use in both sources is not directly comparable.

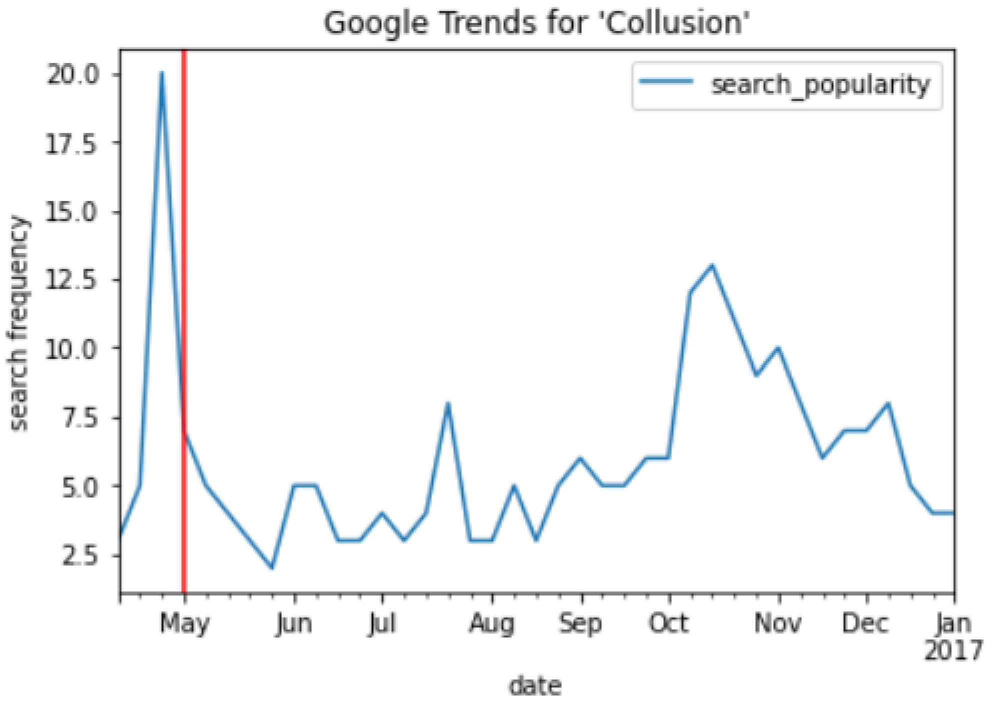


Description and Findings:

The figure above shows the Twitter rankings for ‘China Virus’ within the US in the middle of 2020 graphed as the continuous blue trend line. Trump’s first with the phrase ‘china virus’ is plotted as the vertical red line.

While comparatively newer than the other trending keywords, “China Virus” has been a highly talked about phrase throughout the previous year. Trump’s first tweet appears to fall in May, which seems unusual given the severity of the pandemic as early as March of 2020. This could be due to Trump using this problematic phrase more frequently on live television rather than social media, or in response to other sources using these words to discuss it. A clear connection cannot be drawn between Trump and his influence on this phrase, as the trend is likely too new to have been captured by this particular dataset.

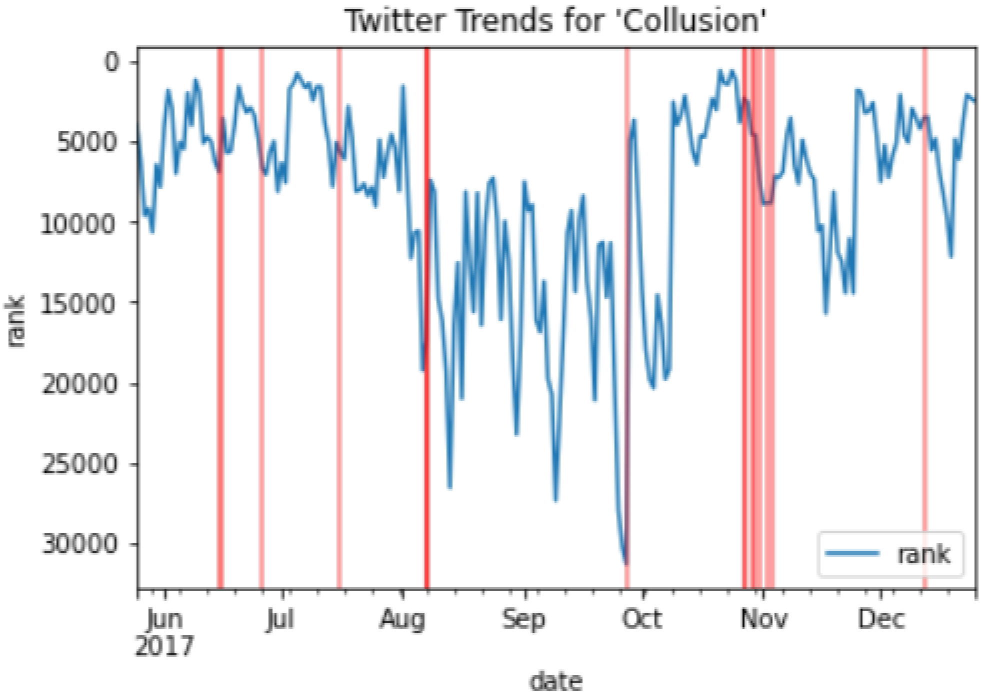
F. ‘Collusion’ Term Analysis



Description and Findings:

The figure above shows the Google trend line for ‘collusion’ searches within the US from mid 2016 to the early 2017 graphed as the continuous blue trend line. Trump’s first tweet with the phrase ‘collusion’ is plotted as the vertical red line.

The dramatic increase in Google searches for this phrase suggests that users were trying to understand the word and its meaning in May of 2016. Trump’s first tweet with ‘collusion’ closely follows this spike, and may lend itself as supporting evidence to Trump reacting to allegations of collusion with and outburst on Twitter.



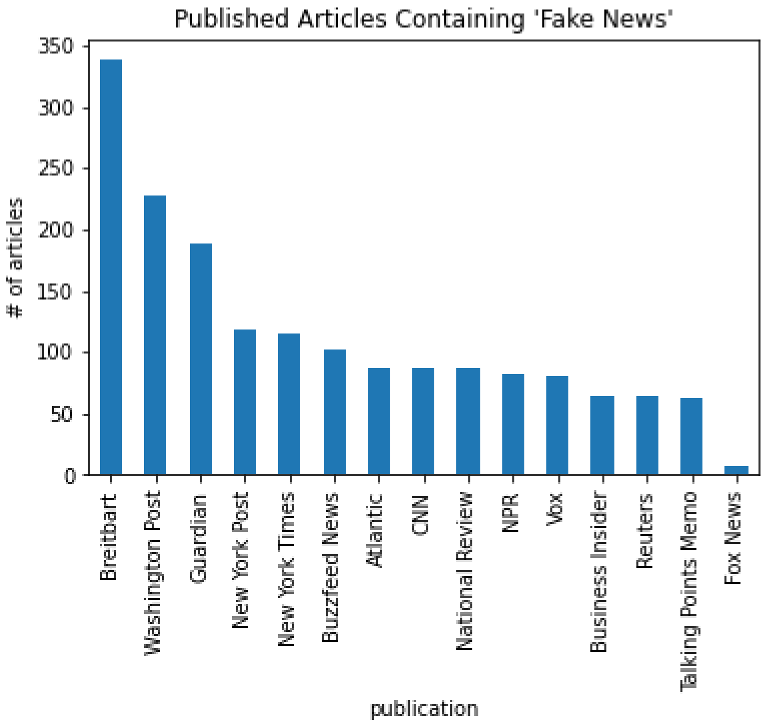
Description and Findings:

The figure above shows the Twitter rankings for ‘collusion’ within the US in the latter half of 2017 as the continuous blue trend line. Trump’s various tweets with this keyword are plotted as the vertical red lines.

Again we see high rankings and popularity of the term ‘collusion’ coinciding with Trump’s various tweets on the matter. He tweeted less frequently about ‘collusion’ during the end of August and through September, which follows the fall in rankings for the term on Twitter overall. This is even more evidence for Trump’s ability to amplify a particular subject at a given time, particularly in response to suggestive jabs at his character and behavior as President.

News Article Analysis

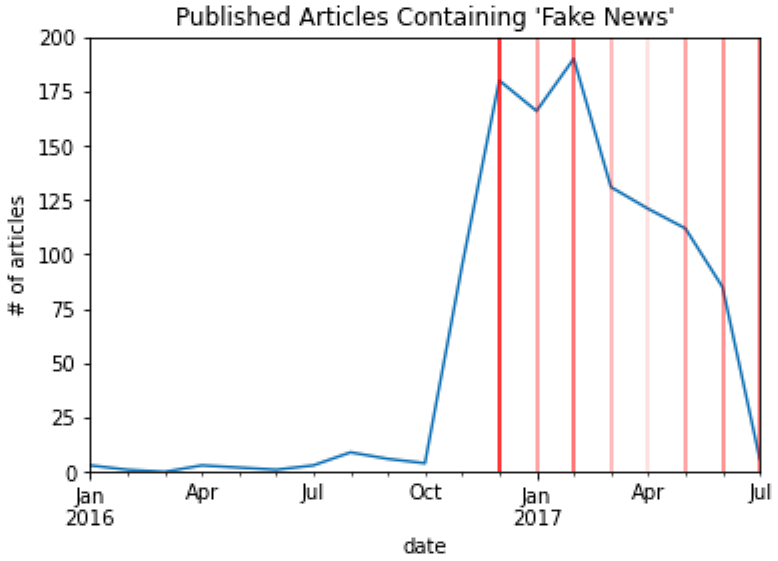
A. ‘Fake News’ Article Analysis



Description and Findings

We returned to our keyword search function to quantify the number of articles that contained the phrase ‘fake news’. Due to the size of the article datasets, we ran the function three times and combined the results into one dataframe, with an article keyword hit per row. We then grouped the data by publisher to display the spread of articles using ‘fake news’ across news outlets.

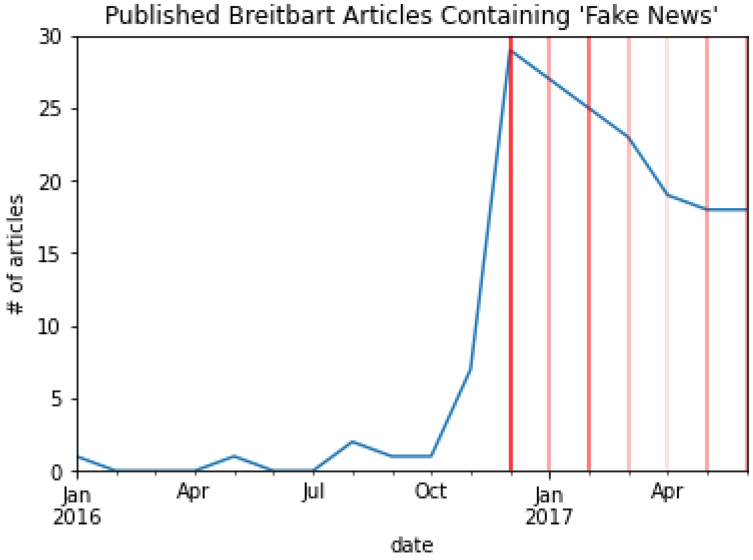
Initially two outliers stand out, Breitbart who published almost 350 articles and Fox News, who published close to 0. There is too little data for us to explore our theories on why Fox News could be so low so we decided to focus on Breitbart's publishing trends and how they related to Donald Trump's tweeting trends in the following figures.



Description and Findings

We referred back to our keyword findings in the previous analysis by using our ‘fake news’ tweets dataframe to compare to the keyword instances across all articles that contain this term. We then plotted the number of ‘fake news’ articles along the blue line, with Trump’s ‘fake news’ tweets overlaid as the red vertical lines on top.

This is the first major piece of evidence that Donald Trump is the primary driving force in terms of media influence. His first tweet (in dark red) is the start of a trend of articles containing "fake news" across many different publications, not just far-right media such as Breitbart. Regardless of which side the publications take, it is clear that Donald Trump has the ability to influence the discussion being held.

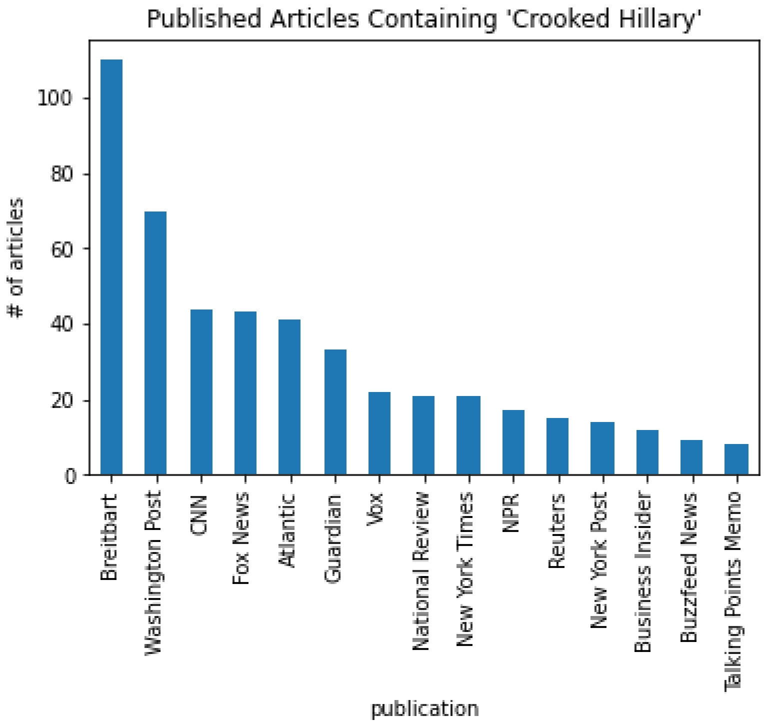


Description and Findings

We again referred back to our keyword findings in the previous analysis by using our ‘fake news’ tweets dataframe to compare to the keyword instances within the most popular publisher. In this case, Breitbart used the phrase ‘fake news’ the most times in our given time period of Jan. 1st 2016 through the end of 2017. We then plotted the number of Breitbart ‘fake news’ articles along the blue line, with Trump’s ‘fake news’ tweets overlaid as the red vertical lines on top.

It is evident that Breitbart's publishing of articles containing "Fake News" skyrocketed starting in November 2016, the month of the election, and reached its peak in December 2016. Trump's tweets containing "Fake News" began around the same time, starting in December and continuing throughout the next several months at a steady pace. Whether Breitbart (a far-right conservative network) began matching Donald Trump's rhetoric or vice versa is unclear, but there is an uncanny link between the two and how the information is disseminated.

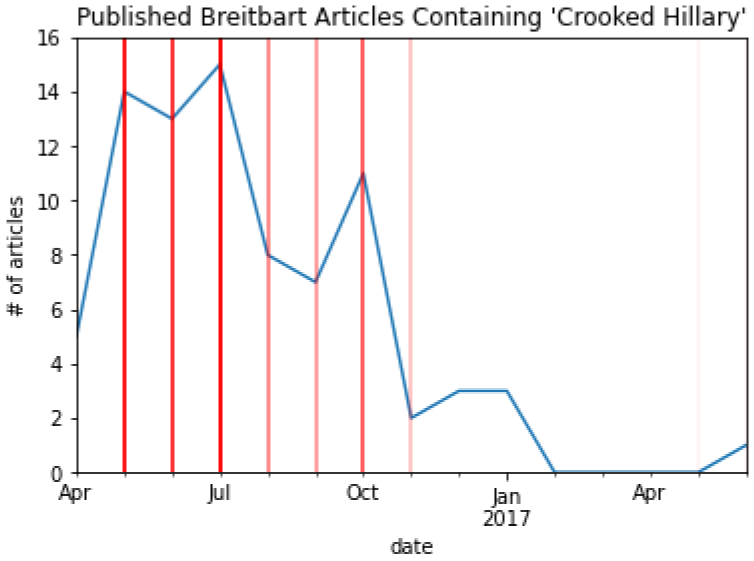
B. ‘Crooked Hillary’ Article Analysis



Description and Findings

This chart displays the number of articles with at least one instance of the term ‘Crooked Hillary’ grouped by publisher to display the spread of articles using this phrase across news outlets.

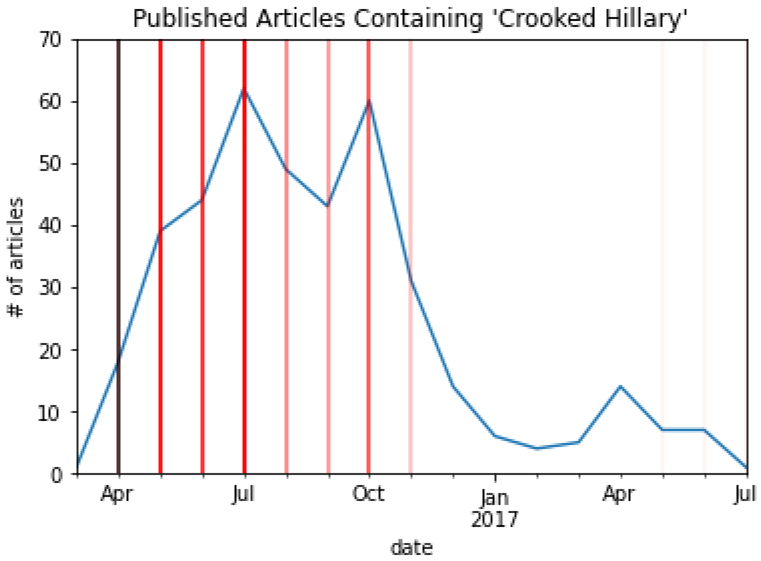
It is clear that Breitbart used the term ‘Crooked Hillary’ more than any other news outlet in this dataset. It is also interesting to note that Fox News is not reflected in the plot at all, likely due to the dataset itself. It would require further exploration to draw meaningful conclusions based on this information.



Description and Findings

Our keyword findings of ‘Crooked Hillary’ in the previous analysis were compared to the keyword instances within the most popular publisher. In this case, Breitbart used the phrase ‘Crooked Hillary’’ more than any other outlet from Jan. 1st 2016 through the end of 2017. We plotted the number of Breitbart ‘Crooked Hillary’ articles along the blue line, with Trump’s ‘Crooked Hillary’ tweets overlaid as the red vertical lines on top.

Here is another example of Breitbart publishing a number of articles containing a phrase around the same time Trump was tweeting about it. Trump had tweeted quite a bit during this time range as indicated by the darker red lines. Additionally, both Trump's tweeting of the phrase and Breitbart's publishing taper off at around the same time.

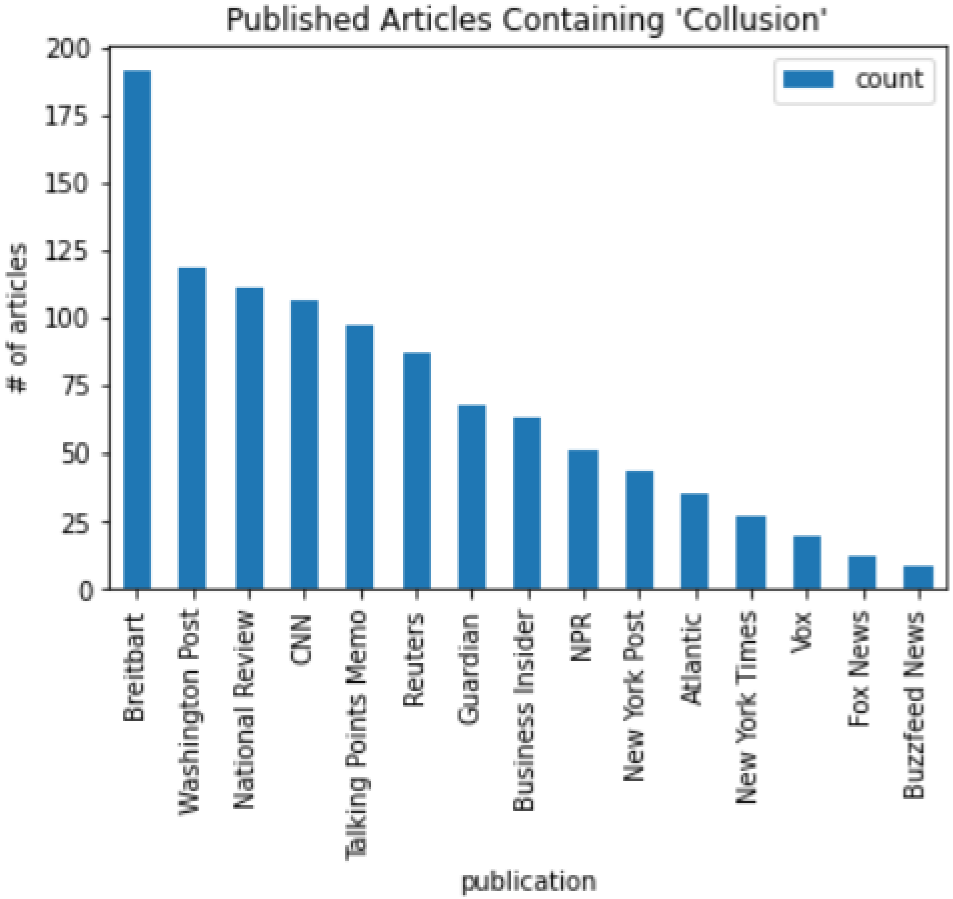


Description and Findings

We referred back to our keyword findings in the previous analysis by using our ‘Crooked Hillary’ tweets dataframe to compare to the keyword instances across all articles that contain this term. All articles that used this phrase from mid 2016 to the summer of 2017 are marked with the blue line. Trump’s ‘Crooked Hillary’ tweets are displayed as the vertical red lines.

Other media outlets don't appear to jump on the trends at quite the same speed as Breitbart, but there is strong evidence that suggests they will cover what Donald Trump tweets. Many of these articles were published prior to Donald Trump’s time in office while Donald Trump was campaigning for President. The term "Crooked Hillary", unlike "Fake News" was not in the common lexicon prior to the 2016 election because it was invented as an insult and a means of undermining the democratic candidate. And our data suggest that the media played into this plan by perpetuation and amplifying the term, because of Donald Trump's continued use of it.

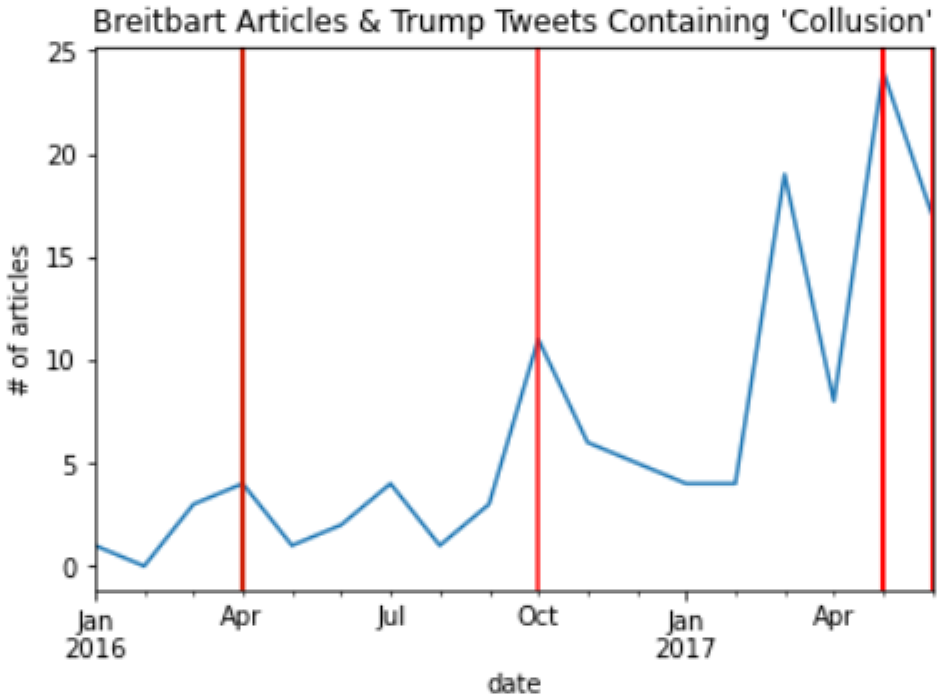
C. ‘Collusion’ Article Analysis



Description and Findings

This chart displays the number of articles with at least one instance of the term ‘collusion’ grouped by publisher to display the spread of articles using this phrase across news outlets.

It is clear that Breitbart used the term ‘collusion’ more than any other news outlet in this dataset. Similar to our ‘Crooked Hillary’ findings, it is interesting to note that Fox News is so low on the article count. This is likely due to the lack of randomized sampling when the dataset was aggregated.

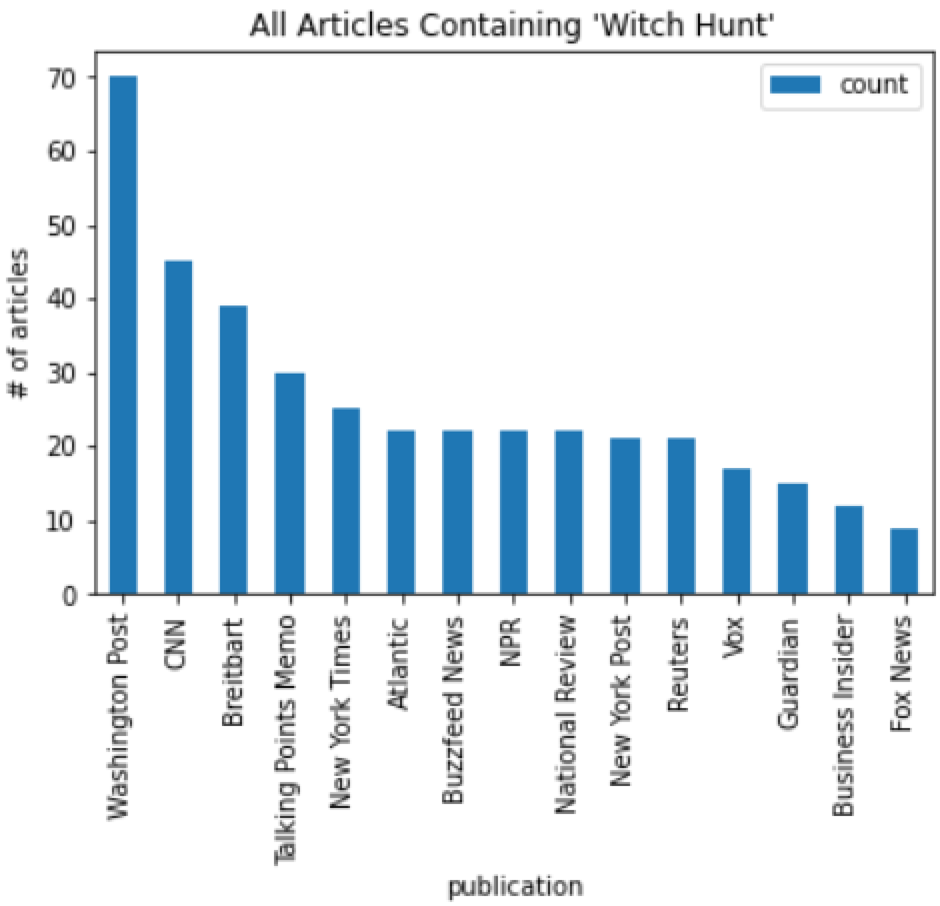


Description and Findings

Breitbart articles that used the term ‘collusion’ from mid 2016 to the summer of 2017 are marked with the blue line. Trump’s ‘collusion’ tweets are displayed as the vertical red lines.

Breitbart’s months with the most articles containing ‘collusion’ coincide with the peaks of Trump’s tweets about the topic. The highest spikes appear in May and June of 2017, when the Senate Judiciary Committee meeting took place in which Comey said he believed Russia was interfering with US politics. He was fired by Trump soon after, and Mueller was appointed as special counsel. While difficult to discern who influenced who, it is clear that there is a relationship between Trump tweeting about his innocence as far as collusion and Breitbart picking up on the story.

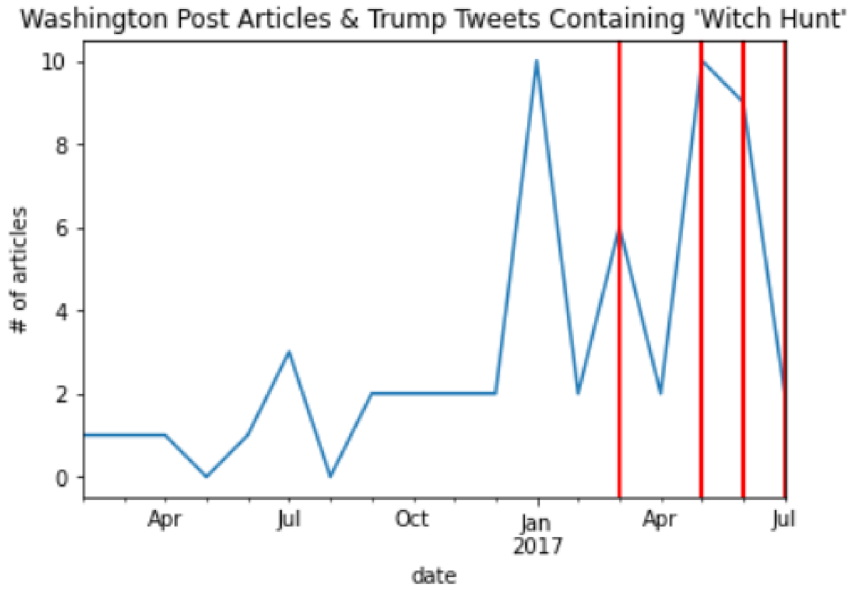
D. ‘Witch Hunt’ Article Analysis



Description and Findings

This chart displays the number of articles with at least one instance of the phrase ‘witch hunt’ grouped by publisher to display the spread of articles using this phrase across news outlets.

It is evident that the Washington Post used the term ‘witch hunt’ more than any other news outlet in this dataset, however the difference in article counts between the other news publishers is not as pronounced as some of the other terms. Perhaps more articles used the phrase ‘witch hunt’ regardless of political affiliation.



Description and Findings

Washington Post articles that used the term ‘witch hunt’ from mid 2016 to the summer of 2017 are marked with the blue line. Trump’s ‘witch hunt’ tweets are displayed as the vertical red lines.

The Washington Post’s months with the most articles containing the key phrase ‘witch hunt’ coincide with the peaks of Trump’s tweets about the topic. An interesting note on this plot is that there was a high number of ‘witch hunt’ Washington Post articles in Jan. 2017, before Trump’s first tweet with the term. It is possible that Trump reacted to the articles on Twitter, in which case the news influenced him rather than the other way around.

Conclusion:

Our analysis of Twitter, Google, and News data helped us answer some of the research questions we set out to explore. The initial keyword analysis helped us grasp just how prevalent certain false information could be across these datasets. We then compared Trump's infamous tweets with Google and Twitter trend data in order to better understand the symbiotic relationship between our former President and the general public's behaviors on two of the most popular websites in the world. While some terms had clear correlations between Trump and ranking or frequency, others did not catch as on as well. As participants in social media, this is inline with our first hand feelings of witnessing Trump tweet these phrases and the adoption of them into mainstream media soon after.

Further analysis between Trump's tweets and news publications revealed that there is indeed a connection between his actions and what is talked about across news sources regardless of political party affiliation. While right-leaning publishers were expected to have more articles containing these fake news phrases, that was not always the case. This is a somewhat harrowing realization and begs the question - which news is real news? While Trump and 'fake news' may be associated as going hand in hand, more exploration would need to be performed to draw stronger connections between Trump and the media.