Nasty, Brutish, and Short: What Makes Election News Popular on Twitter?

Anonymous

Abstract

AAAI creates proceedings, working notes, and technical reports directly from electronic source furnished by the authors. To ensure that all papers in the publication have a uniform appearance, authors must adhere to the following instructions.

Introduction

In the changing landscape of both journalism and politics, social media is playing an increasingly large role in mobilizing and spreading information to citizens. A Pew Research survey from August 2015 showed that nearly two-thirds of adults in the U.S. who are on Twitter use the platform to get news (Bartherl et al. 2015). During the 2016 election year, the New York Times published an article estimating a 2 billion-dollar advantage in free media, including social media, for Donald Trump, all of which has no small impact on the messages broadcast to voters [13].

Although social media messages are less able to be carefully controlled in comparison to paid advertisements, they also have the potential to reach a wider audience. Reactions to an article shared by one potential voter now have the ability to be broadcast and spread to millions of others in a real-time, public sphere.

The popularity of sharing articles on social media also marks an important shift in the role of the news consumer from armchair reader to information propagator. Whereas news used to be broadcast to the reader, now each reader has the potential to broadcast stories to his or her own audience. Sharing a story requires a level of interest and activation on the part of the reader beyond simply reading a story; yet often, this trigger is predictably emotional in nature.

In a 2011 study of the New York Times "most emailed list", Berger and Milkman found that the potential for a news story to go viral is partially driven by physiological arousal, defined as "an excitatory state of sensory alertness, mobilization, or energy" [Milkman 7]. In short, the desire to share a certain story is often universally impulsive, regardless of context. Yet in the case of political news, this impulse can have a large impact on reach of political messages.

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Hypotheses

With the impact of social media in politics *and* the often emotionally charged nature of content-sharing in mind, we ask the following question in our research:

- Does the emotional vocabulary of political news stories have an impact on its Twitter popularity that persists beyond political affiliation?

To test this question, we focus on three key aspects of stories: length, emotionality, and positivity, based on behavioral theories of the Internet and studies of political news detailed in the section below.

We hypothesize the following behavior in our dataset of stories and tweets:

- **H1:** Story length has a *negative* correlation with Twitter shares, due to the effects of the Internet attention economy and overexposure to political media (Goldhaber 1997).
- **H2:** Emotionality has a *positive* correlation with Twitter shares, consistent for viral content in general (Berger and Milkman 2012).
- **H3:** Positivity has a *negative* correlation with Twitter shares, due to the nature of political news and contrary to generalized findings (Berger and Milkman 2012)

For each of these three independent variables (story length, emotionality, positivity) we repeat analyses across three views of the data: first, the entire dataset; then, by political candidate followed amongst users who follow only one candidate; and finally, by the number of political candidates followed (degree of political engagement), to look for differences amongst different populations of political tweeters.

Literature Review

The (Short) Attention Economy

At the same time that social media has the power to create a flood of free advertising and media for political candidates, the abundance of information on the web has created new challenges and questions about the kind of content being processed by readers. This paradox– between the ease of accessibility to information and the increasingly limited bandwidth of consumers– is described as one of the challenges of being in an *attention economy* (Goldhaber 1997).

Moreover, high-impact events like the presidential elections especially intensifies this effect– about 60 % of Americans reported feeling exhausted by media coverage of the elections in July of 2016 (Gottfried 2016). To explore the effects of the attention economy on the reading of political news, we examine story length and how it relates to sharing popularity in the analysis to follow.

Negativity in Politics and the Internet

In addition, the option of anonymity and pseudo-anonymity on a social network like Twitter (along with other traits of Internet communication), is theorized to contribute to increased negative and hostile behavior, potentially increasing tension for the already-fraught subject of politics. This phenomenon, is coined as the *online disinhibition effect* (Suler 2004).

In Berger and Milkmans study of story virality, it was found that *positive* content was more likely to be shared than negative content– against conventional belief (Berger and Milkman 2012). Political news, however, is a unique category of news, and this election in particular– where one-in-four Americans report disliking the presidential candidates– appears to have a negative overtone.

To compare the sharing of election news stories versus patterns of general virality in the news, and to examine the extent in which negative sentiment is popular, we calculate the *negativity* of stories, and how that relates to Twitter behavior.

We also examine the effects of the degree of combined emotionality in the content and how that relates to Twitter shares, to see if either more positive or more negative content is more likely to be shared overall than content that ranks low in emotionality. Although positive content was found to be more popular than negative content in the sharing of stories, both highly positive and highly negative content was more likely to become viral, and we expect the same to hold for political news (Berger and Milkman 2012).

Data Collection

Our main dataset is a connected corpus of news articles about the presidential elections and the tweets that share them from January 1, 2016 to May 1, 2016 over 13 news outlets. In this chapter, we detail the collection and creation of our dataset, along with the unique challenges of working with "big data".

The Electome Project

The backbone of our data collection and classification process lies under the umbrella of the Electome project, a large, collaborative effort with the Laboratory for Social Machines to examine the "horse-race of ideas" and competition of narratives during the presidential election year. The goal of the Electome project is to create novel stories for data journalists as well as technically innovative ways to examine the national conversation using machine learning and "big data" analysis through both social and traditional media coverage. The first step of both our news and Twitter data processing method uses machine learning classifiers from the Electome project.

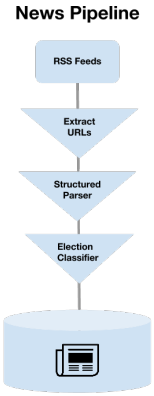

Figure 1: Election News Pipeline

News Dataset

Duration & Scope For our news dataset, we scraped articles from the RSS feeds of news publications every hour over five months and 13 publications:

- CNN
- Fox News
- The New York Times
- The Wall Street Journal
- The Washington Post
- The Los Angeles Times
- The Associated Press
- Reuters
- McClatchy
- Politico
- Buzzfeed
- The Huffington Post
- NPR

The choices above span a mix of publications. We include sources that:

- Have mostly conservative audiences and mostly liberal audiences (**?**)
- Come from mixed primary media formats (television, paper, online, radio)
- Are viewed as "legacy" (over a hundred years old) and "new" media (founded online within the last 10 years)
- Focus solely on political news (Politico, McClatchy)
- Are newswire services (the Associated Press, Reuters news)

to capture a variety of types of election coverage and target audiences.

We look at stories from January 1, 2016 (the start of the election year) to May 1, 2016. This time period captures the bulk of the primary election, when coverage of multiple presidential candidate contenders creates greater variety in news stories for our analysis.

Data Pipeline

Articles are processed in a 3-step pipeline, pictured below.

After collecting the links to the full content of the news stories from each publication's RSS feed, we pass each link

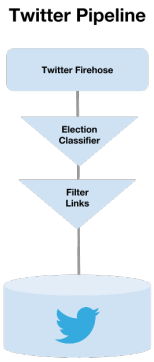

Figure 2: Election Tweets Pipeline

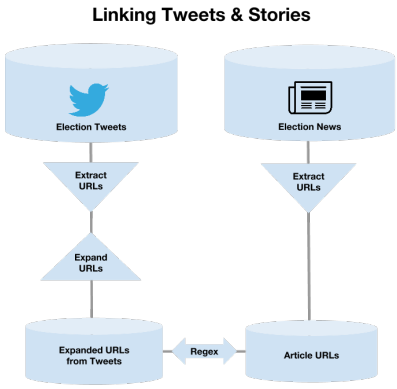

Figure 3: Data Pipeline

to a structured content parser that extracts entities and features from the raw HTML.

The story text is then passed into a machine learning classifier for election news from the Electome project [1].

We collect and classify a total of 22,959 articles as election-related with over 80% confidence, an average of 5,700 per month and 191 per day.

Tweets Dataset

Duration & Scope

The Laboratory for Social Machines was founded on the premise of a research grant with Twitter which allows access to the "firehose" of all activity on Twitter. We start with the pool of all tweets between January 1, 2016 and May 1, 2016 in our data process.

Data Pipeline

Tweets pass through a similiar pipeline as news stories. We sort all tweets with an election classifier which has been shown to be able to detect election-related tweets with an F-score of 92% (**?**). We then filter by those that share a link (which might potentially be a news story).

We collect and sort a total of 16,667,685 tweets as election-related and containing at least one URL in the text, an average of 4,000,000 per month and 140,000 per day.

Combined Dataset

The final step of our data collection process is to extract, expand and connect the links shared in our election-related tweets with articles in our database.

Mapping Tweets to Stories

Twitter automatically formats all links into a shortened "t.co" format, so we first expand all links in tweets (16.6 million), then use a regular expression to see if the final destination of the expanded link matches a query-truncated URL of a story in our database. We checked the validity of

[1] Designed and implemented by Prashanth Vijayaraghavan.

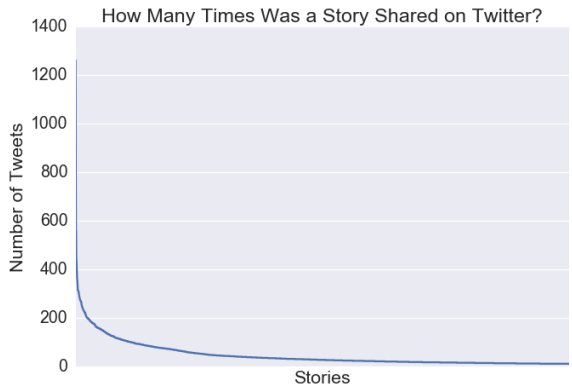

Figure 4: Distribution of Story Shares

382 billion url-story matches in less than a day by running the processes on the Amazon Web Services cloud computing platform in parallel using the Gnu-parallel command line tool (**?**).

Final Corpus

In total, we found that 30% of the election stories we tracked were shared on Twitter during the time period of January 1st through May 1st.

There were 137,986 tweets that contained a link to 6,911 unique stories (out of 22,960).

Since we chose the story to be the unit of analysis in this thesis, we then eliminated any stories that were shared by less than 10 tweets.

This left a total of 2,650 distinct articles (38%) shared in 123,113 (89%) tweets by 20,956 Twitter users (93%).

Descriptive Findings

The vast majority of stories are shared less than 100 times.

Story sharing behavior follows an approximate power law distribution. On average, stories are shared 46 times, however, the median (50th percentile) of shares is just 26.

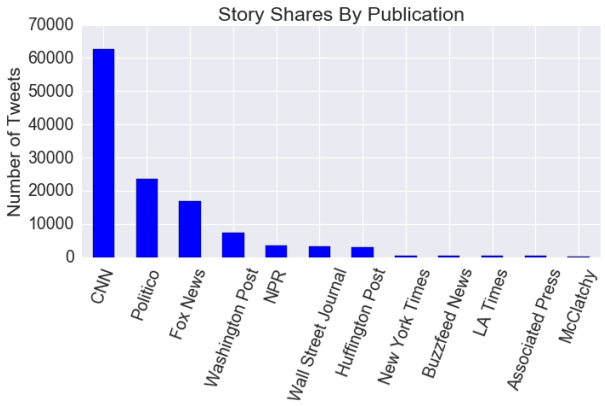

Figure 5: Number of story shares by publication

CNN, Politico, and Fox lead in publication popularity with the highest number of stories shared by tweets in our dataset– likely due to the volume and close association to political content of the companies. Because our data pipeline detailed in the previous chapter looks for election-related news, outlets which track the campaign extensively are more likely to show up in our results.

Article
The One Weird Trait That Predicts Whether You're a Trump Supporter
Donald Trump Is Shocking, Vulgar and Right
Biden praises Sanders on income inequality
Why I'm voting for Trump
Anne Frank's stepsister compares Donald Trump to Adolf Hitler
Trump basks in his spotlight
Rubio: Law-abiding undocumented immigrants could stay
Terrorists use Trump's 'Muslim ban' speech in recruitment video
Iowa caucuses: Donald Trump's moment of truth
GOP senators: If Cruz wins, we lose

Table 1: Top 10 Shared Stories

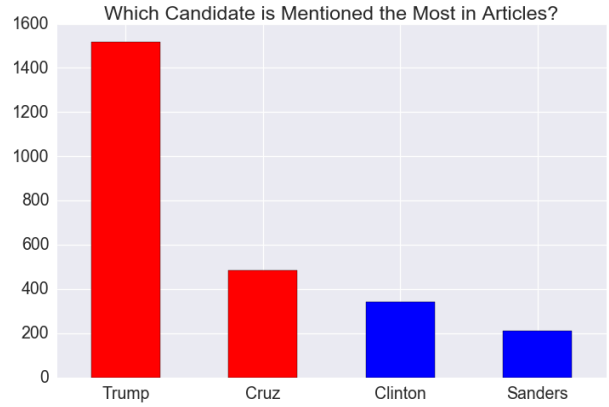

Figure 6: Most frequently mentioned candidate in stories

Examining the top 10 most shared stories, it comes as no surprise that outsized personality Donald Trump is by far the most "tweetable" candidate, dominating the list with 7 out of 10 stories featuring his name in the title.

The *extent* to which he is prominent in articles is clear: by coding each story by the most frequently mentioned candidate, Trump has nearly three times as much coverage at nearly 60% as the runner-ups, Ted Cruz and Hillary Clinton.

The large number of stories with Cruz as the most-mentioned candidate are likely due to his association to Trump as a Republican runner-up: 96% of stories where Cruz is the most-mentioned candidate feature Trump as the second-most frequently occuring.

References

[Bartherl et al. 2015] Bartherl, M.; Shearer, E.; Gottfried, J.; and Mitchell, A. 2015. How do americans use twitter for news? Retrieved from `http://www.pewresearch.org`. (Accessed on 07/12/2016).

[Berger and Milkman 2012] Berger, J., and Milkman, K. L. 2012. What makes online content viral? *Journal of marketing research* 49(2):192–205.

[Goldhaber 1997] Goldhaber, M. H. 1997. The attention economy and the net. *First Monday* 2(4).

[Gottfried 2016] Gottfried, J. 2016. Most americans already

feel election coverage fatigue. Retrieved from http://www.pewresearch.org. (Accessed on 07/16/2016).

[Suler 2004] Suler, J. 2004. The online disinhibition effect. *Cyberpsychology & behavior* 7(3):321–326.