Higher Ground? How Groundtruth Labeling Impacts Our Understanding of the Spread of Fake News During the 2016 Election

Abstract

The spread of fake news in online social media platforms has garnered much public attention and apprehension. Consequently, both the industry and academia alike are investing increased effort to understand, detect, and curb fake news. Yet, researchers differ in what they consider to be fake news sites. In this paper, we first aggregate 5 distinct lists of fake news sites, and 3 lists of mainstream news sites published by experts and reputable organizations. Then, using each pair of fake and mainstream news lists as an independent "groundtruth", we examine i) the prevalence and ii) temporal characteristics of fake news as well as iii) the agenda-setting differences between fake and mainstream news sites. We observe that depending on the groundtruth, the prevalence of fake news varies significantly. However, the temporal trends and agenda-setting differences between fake and mainstream news sites remain moderately consistent across different groundtruth lists.

1 Introduction

Following the 2016 U.S. presidential election, fake news swiftly became a topic of interest, scrutiny, and apprehension for political pundits, media scholars, and the general public alike. Many worry that fake news is weakening public trust in both established political institutions and mainstream news media (Silverman, 2017; Guo and Vargo, 2018). These concerns are driving increased research efforts on fake news from academia and industry alike. One of the most significant challenges for the research community has been how to define fake news. While there is currently no consensus on the topic, leading scholars advocate"... focusing on the original sources—the publishers—rather than individual stories, because we view the defining element of fake news to be the intent and processes of the publisher." (Lazer et al., 2018). Yet, there is currently no agreement on which news producers are fake news producers either (Tandoc Jr, Lim, and Ling, 2018).

Consequently, there are a number of lists with opaque generation processes (Zimdars, 2016; Guo and Vargo, 2018; Allcott and Gentzkow, 2017; Van Zandt, Dave, 2018; Politifact staff, 2018; Shao et al., 2016) being used by studies with important implications such as examining fake news cascading behavior (Allcott and Gentzkow, 2017; Allcott, Gentzkow, and Yu, 2018), assessing agenda setting powers of fake and traditional news sites (Vargo, Guo, and Amazeen, 2018; Guo and Vargo, 2018; Mukerji, 2018) or characterizing changes in fake news trends (Allcott, Gentzkow, and Yu, 2018). How robust are these studies with respect to the choice of groundtruth lists that define which publishers are producers or fake and traditional news? We aim to answer this question in this paper.

We aggregate 5 lists of fake news sites and 3 lists of mainstream news sites contributed by both the academia and other reputable sources (Poynter Institute, 2019; Zimdars, 2016; Wang, 2017; White, Nicholas, 2018; Leetaru and Schrodt, 2013; Van Zandt, Dave, 2018). We first review the labeling processes of these lists, measure their similarities, and assess their temporal changes. We then determine how groundtruth selection impact downstream observations from 3 distinct dimensions.

We first examine prevalence given the divergent findings in recent work (Silverman, 2016; Allcott and Gentzkow, 2017; Bovet and Makse, 2019; Grinberg et al., 2018). For instance, prior work by Silverman (2017) shows that selective fake news articles "garnered hundreds of thousands of shares – sometimes more than twice as many as legitimate news scoops in major outlets". Whereas, a study by Allcott and Gentkow (Allcott and Gentzkow, 2017) suggests that an average adult only saw and remembered 1.14 fake news article during the course of the 2016 election. Next, we investigate the robustness of trend analysis since having an accurate assessment of temporal patterns of fake news such as trend can assist platform owners in evaluating whether their efforts to curtail fake news is successful (Allcott, Gentzkow, and Yu, 2018). Finally, we turn to topic analysis. Agenda-setting theory (McCombs and Shaw, 1972) postulates that the most frequently covered topics are what the general public considers the most important. Within the context of misinformation and the 2016 Election, fake news sites may have led voters to re-evaluate issue importance and candidate viability by prioritizing certain topics over others (Guo and Vargo, 2018; Vargo, Guo, and Amazeen, 2018).

Our paper makes the following contributions:

- We demonstrate that existing fake news lists share very few domains in common. Additionally, popular fake news sites are more likely to be included (and included earlier) than unpopular ones. Further, domains that belong to subcategories *hate*, *junksci*, *clickbait* in Zimdars's list (2016) are less likely to be included by other lists compared to domains in the *fake* subcategory.
- Based on the groundtruth choice, the prevalence of fake news varies considerably (2-to-40%). This discrepancy is mostly due to whether a list includes domains with mixed factualness.
- We show that the time series correlation between most lists is high, especially for the general election period. In fact, we observe an increase in fake news prevalence during the general election regardless of groundtruth choice. However, results are less consistent for primary and after election periods. Further, we also show that scheduled events contribute to a temporary drop in fake news prevalence. Observations for scandals are not as robust and are dependent on groundtruth selection.
- Studying the agenda-setting priority difference between fake and traditional news sites, we observe that whether a topic (e.g. immigration) was more central to the coverage of fake news outlets compared to the traditional news sites is robust to the choice of groundtruth.
- Finally, groundtruth selection of mainstream news list has very limited impact on all 3 aforementioned downstream analysis.

To summarize, we argue that despite fake news annotators have varied labeling and validation procedures as well as have very few domains in common, the domains within each list, however, are comparable to each other and demonstrate similar characteristic (e.g. agenda-setting). Thus, while groundtruth selection affects prevalence analysis, it has limited to moderate impact on studies concerning the behaviors of fake news sites.

2 Related Work

Researchers have extensively documented the negative impact fake news has on the quality of civic engagement, healthcare, stock market, and disaster management (Rapoza, 2017; Marcon, Murdoch, and Caulfield, 2017; Palen and Hughes, 2018), both within the United States (Silverman, 2016; Main, 2018; Starbird, 2017) and internationally (Kucharski, 2016; Alimonti and Veridiana, 2018). This is due to individuals' inability to tell false information apart from the truth (Wineburg, 2016), and susceptible to the influence of false information (Balmas, 2014; Pennycook, Cannon, and Rand, 2018).

Many studies aim to distinguish false content from credible news articles at scale. Prior studies have identified differences in i) linguistic patterns such as punctuations and word choices (Potthast et al., 2017), ii) auxiliary data associated with the articles such as the place of publishing (Shu et al., 2017; Shu, Wang, and Liu, 2018), iii) network cascading attributes such as depth, breadth, and speed (Shao et al., 2017; Allcott, Gentzkow, and Yu, 2018; Vosoughi, Roy, and Aral, 2018), and iv) agenda-setting priorities (Vargo, Guo, and Amazeen, 2018). Some differences between fake and credible articles are then used to build automated fake news detection platforms (Horne and Adali, 2017; Horne et al., 2018) in an effort to curtail its spread.

However, efforts to study fake news and to diminish its influence are difficult, partly because scholars do not have a consistent definition for fake news (Tandoc Jr, Lim, and Ling, 2018; Wardle, 2017). For instance, Tandoc et al. identify 2 primary dimensions of fake news: levels of facticity and deception. Wardle, on the other hand, conceptualizes fake news using 3 distinct dimensions: type of content, motivation, and dissemination method. Moreover, existing fake news labelsets (Politifact staff, 2018; Zimdars, 2016; Van Zandt, Dave, 2018; White, Nicholas, 2018; Mitra and Gilbert, 2015; Leetaru and Schrodt, 2013) have considerably different annotation and categorization procedures.

We first consolidate existing groundtruth labelsets of fake and mainstream news sites that have been generated by various groups. We then assess whether and to what extent differences in groundtruth selection affect downstream studies.

3 Data

We use 3 types of data: i) lists of fake and traditional news sites, ii) tweets about the two candidates during the 2016 U.S. Presidential election, and iii) webpages, or news articles, corresponding to the URLs shared in those tweets.

Fake and traditional news Site Lists: We collect 5 distinct fake news lists and 3 traditional news list from both the academia and the press (Zimdars, 2016; Guo and Vargo, 2018; Allcott and Gentzkow, 2017; Van Zandt, Dave, 2018; Politifact staff, 2018; Shao et al., 2016), resulting in 1884 aggregated fake news sites, 8238 traditional news sites, and a combined total of 10K+ news sites. We describe and evaluate these lists in Section 4.

Twitter Data: The social media dataset used in this study is described in detail in Bode et al. (Forthcoming). The data collection was performed using Sysomos MAP - a social media search engine that includes access to all tweets (Twitter firehose) going back one year as well as other social media sites and online news. For any given day between May 23, 2014, and January 1, 2017, our dataset includes i.) 5,000 tweets randomly sampled from all tweets that included the keyword Trump and also contained URLs, and ii) 5,000 tweets similarly sampled from all that mentioned Clinton and had URLs. The resulting dataset includes approximately 4.8 million tweets each about Donald Trump and Hillary Clinton respectively.

Webpages (News Articles): The webpages dataset (Budak, 2019) includes the content of the webpages shared in the Twitter dataset described above. For each tweet with an external URL, the dataset includes a record with: i) the shortened URL, ii) the original URL, iii) domain name, iv) title of the document, v) body of the document, (vi) the date of the tweet, vii) Twitter account id of the user sharing the URL, and viii) a binary categorization that indicates whether this tweet is about Clinton or Trump. We remove the records with domains not listed in the aforementioned 10K+ news sites and filter out the tweets posted before 12/01/2015 or after 01/01/2017. . We derive approximately 244K unique articles shared by 1M Tweets on Twitter.

4 Meta-review

In this section, we first examine the characteristics and applications of the available lists of fake and traditional news websites. Then, focusing on fake news lists, we assess their commonalities and differences and explore the characteristics of websites that are correlated with them being included in or excluded from any given list.

Lists of Fake News Sites: We collect 5 fake news lists.

1. `ZDR`: We refer to the set of fake news websites annotated by Zimdars et al. (2016) as `ZDR`. `ZDR` also further tags each website with at most 3 of out the following 10 subcategories: *fake, satire, bias, conspiracy, rumor, state, junksci, hate, clickbait*, and *unreliable* [1]. Amongst these subcategories, *unreliable* and *clickbait* are noted to have "mixed" factualness.
2. `MBFC`: The set of sites labeled by *Media Bias/Fact Check*– an independent online media outlet maintained by a small team of researchers and journalists (Van Zandt, Dave, 2018)–will be referred to as `MBFC`. Similar to `ZDR`, `MBFC` assigns domains to subcategories: *fake, conspiracy, satire*. Moreover, it also labels websites with ideology (*extreme left, left, center, right, extreme right, unlabeled*) and rates websites by their factualness (*low, mixed, high*).
3. `POLIT`: The staff of PolitiFact, in collaboration with Facebook, identified the list of most-shared fake news sites on Facebook during the 2016 election (Politifact staff, 2018). This list–referred to as `POLIT`–assigns sites to *fake, imposter, some fake*, or *parody*.
4. `DDOT`: This list is shared by *the Daily Dot*, a mainstream online news site (White, Nicholas, 2018). This list is largely created by referencing other pre-existing fake news lists and does not contain subcategories.
5. `AGZ`: Allcott et al. (2018) aggregated the following five lists into one: `POLIT`, Grinberg et al. (2018), Silverman (2016), Schaedel (2018), and Guess et al. (2018). This list is referred to as `AGZ`. The subcategorization process in `AGZ` is somewhat complex. For instance, while `POLIT` was one of the lists considered by `AGZ`, its subcategories were ignored and all the domains were re-labeled as *fake*. However, the subcategories *black, red, orange* (black being completely false, and red/orange being with unreliable claims) of Grinberg et al. (2018) were maintained. Finally, all domains from other referenced lists were labeled as *fake*.

A synthesis of these lists reveals that 4 out of the 5 lists share 2 common subcategories: i) a subcategory containing domains with *mixed* factualness, and ii) a *fake* subcategory (entirely fabricated information). This consistency suggests that domains belong to *mixed* or *fake* are conceptually distinct from others. Thus, studies on fake news should take this distinction into consideration.

Lists of Traditional News Sites: We consider the following three traditional news lists.

1. `ALEXA`: Alexa is an online domain directory owned by Amazon (Wikipedia contributors, 2019). We crawl for all the websites listed under Alexa's *News* category.
2. `MBFC(T)`: *Media Bias/Fact Check* also lists a large set of traditional news sites. We refer to this list as `MBFC(T)`.
3. `VARGO`: This list contains fact-based news websites compiled through manual content analysis of the top news media websites found in GDELT 's global knowledge graph (Vargo, Guo, and Amazeen, 2018).

For the fake news lists, `DDOT` has the smallest size of 175 domains, followed by `POLIT` and `AGZ` at 327 and 673, `ZDR` and `MBFC` are the largest lists with sizes of 786 and 1183. In addition, list sizes are 1685, 2649, and 5497 for `MBFC(T)`, `VARGO` and `ALEXA`. Table 1 provides a summary of the annotation processes and the uses of these lists. As is evident from the second column (*Annotation and Quality*), most lists do not have a transparent annotation and quality evaluation procedure. Perhaps due to the absence of such robust procedures, there is no consensus on which of these lists should be treated as the ultimate ground truth. This is clear from the third column (*Applications*). More than 20 studies have used these lists of fake and traditional news sites. The lists are used for various important purposes such as building automated fake news classifiers or assessing fake news' impact on the 2016 election. This highlights the importance of identifying similarities and differences between the lists.

Thus, we conduct downstream analysis using different groundtruth pairs (f,t) where $f \in \{\texttt{ZDR},\texttt{MBFC},\texttt{POLIT},\texttt{DDOT},\texttt{AGZ}\}$, and $t \in \{\texttt{ALEXA},\texttt{MBFC(T)},\texttt{VARGO}\}$.

List Overlap: Here, we identify the overlap among the 5 fake news lists using 2 metrics. We first calculate the fraction of websites being present in at least 2 of the 5 lists, then 3, then 4. We observe that close to 50% of all domains are only included in a single list. In fact, only 5.7% of the domains are included by all fake lists. Second, we also calculate the Jaccard similarity score (Goodall, 1966) of each pair of lists. We observe that more than half of the pairs of fake news lists have a similarity of $<= 0.1$. We note that `MBFC` and `DDOT` have the lowest Jaccard similarity score of 0.08, and `AGZ` and `POLIT` have the highest score of 0.48.

The extent of dissimilarity between the lists is surprising and we identify four potential measures here: i) *popularity*, defined as the number of times a URL from a given domain is shared in the Twitter dataset, ii) *age* (we collect data using whois.com, an online domain registration service), iii) *subcategory*, as defined by Zimdars et al. (2016)[2], and finally

[1] Zimdars et al. also list a small subset of domains as *political, reliable* and *unidentified* which are not fake news sites and therefore removed from subsequent analyses.

[2] Zimdars et al. have the most comprehensive subcategories and a coherent labeling guideline. Subcategory is *unknown* if a domain is not listed by Zimdars et al. (2016).

Table 1: Traditional and Fake News Lists and Their Applications. Some studies below use multiple sources.

List	Annotation and Quality	Applications
DDOT	no information	build automated fake news trackers (Shao et al., 2016; Helmstetter and Paulheim, 2018), assess agenda setting powers of fake and traditional news sites (Vargo, Guo, and Amazeen, 2018; Guo and Vargo, 2018; Mukerji, 2018)
AGZ	authors aggregate lists generated by others, and then use various combinations of these list for result robustness check	assess impact on election, examine fake news cascading behavior (Allcott and Gentzkow, 2017); examining fake news trend (Allcott, Gentzkow, and Yu, 2018)
MBFC	annotated by staff; authors examine wording, source, story selection, and political affiliation	studies of the Alt-right (Main, 2018), globalism (Starbird, 2017), the virality of fake news (Darwish, Magdy, and Zanouda, 2017), information literacy (Farmer, 2017), polarization (Croft and Moore, 2017), and information quality (Nelimarkka, Laaksonen, and Semaan, 2018)
POLIT	no information	study the diffusion of fake news on social media (Allcott, Gentzkow, and Yu, 2018), information literacy (Mukerji, 2018), automate fake news detection (Granskogen, 2018)
ZDR	annotated by scholars and librarians; domain name, about us page, writing style, aesthetics, and social media accounts are among the examined characteristics	examine network cascading behavior difference between fake and real news articles during the 2016 Election (Allcott and Gentzkow, 2017; Allcott, Gentzkow, and Yu, 2018), build fake news classifiers (Shao et al., 2016; Horne and Adali, 2017; Horne et al., 2018), assess agenda setting powers of fake and real news sites (Vargo, Guo, and Amazeen, 2018; Guo and Vargo, 2018), impact assessment (Rini, 2017; Figueira and Oliveira, 2017; Doshi et al., 2018), ethics and policy (Farte and Obada, 2018; Koulolias et al., 2018)
MBFC-T	see MBFC	see MBFC
ALEXA	no information	examine cascading behavior differences between fake and traditional news articles (Allcott and Gentzkow, 2017; Allcott, Gentzkow, and Yu, 2018), news sharing behavior in right-leaning echo chambers (Lima et al., 2018)
VARGO	annotated by authors; intercoder reliability of 0.988 Krippendorff's alpha.	assess agenda-setting power of fake and real news sites (Guo and Vargo, 2018; Vargo, Guo, and Amazeen, 2018)

vi) *ideology*, as defined by *Media Bias/Fact Check*(2018)[3]

The details of the regression model and the analysis are provided in the Appendix. We observe that the popularity of a website is positively correlated with being included in lists (though the variable is not significant for DDOT and POLIT). Further, ideology is not predictive of whether a website is more likely to be included by lists except for AGZ (conservative-leaning domains are more likely to be listed). Finally, we observe that compared to domains subcategorized as *fake* by ZDR, domains belong to other subcategories are uniformly less likely to be present in all the other lists.

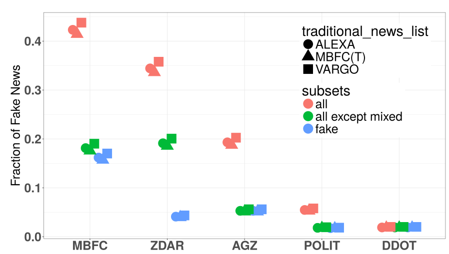
We further examined how the lists changed over time and found the types of changes to be largely consistent. For the lists we have temporal information for (MBFC, ZDR, and DDOT), we observe the following: (i) they include more popular domains earlier on–adding the less popular ones later, (ii.) they include the sites that publish fake news earlier compared to sites that publish less problematic categories such as clickbait, and (iii.) interestingly, sites labeled as *liberal* and *conservative* are comparable in when they are added to the aforementioned lists. For the regression model for temporal analysis, we refer the reader to the Appendix.

5 GroundTruth Selection and Downstream Consequences

A meta-review of the fake news lists in the previous section demonstrates marked differences between these lists. How do these differences affect the downstream analysis? We aim to answer this question in this section. To that end, we first assess how groundtruth selection impact the perceived prevalence of fake news during the 2016 election. Next, we measure the similarities or dissimilarities of fake news time-series generated using different groundtruth pairs (f,t). Finally, we determine whether there are any marked differences in agenda-setting priorities of fake and real news sites due to choice in groundtruth.

[3]Ideology is *unknown* if the domain is not listed by *Media Bias/Fact Check*(2018) or if *Media Bias/Fact Check* didn't mark it with an ideological label (approximately 18.6% domains). Here we collapse MBFC's *extreme left* and *left* categories into single *liberal* class. Same for *conservative*.

Figure 1: Fraction of Fake News. The x-axis indicates fake news lists. Each list is divided into subsets (marked by color) of *all*, *all-except-mixed* (not including domains in *mixed* subcategory), and *fake* (only domains in *fake* subcategory). The shape of each point denotes mainstream news lists, and y-axis is the fraction of tweets contain fake news.



5.1 Prevalence

Here, we define *prevalence* as the fraction of tweets containing URLs that are from fake news sites. We examine to what extent groundtruth difference impacts perceived pervasiveness of fake news using 3 distinct boundary conditions (strictness in definition) for each fake news list: *all*, *all-except-mixed*, and *fake*. More specifically, given a groundtruth pair (f,t), we write f_{all} as the entire set of domains in f, f_{mixed} and f_{fake} as the set of domains in f_{all} that belong to subcategories with mixed factualness and the subcategory *fake* respectively. We then calculate *prevalence* as

$\frac{|f_{all}|^s}{|f_{all}|^s+|t_{all}|^s}$, $\frac{|f_{all}|^s-|f_{mixed}|^s}{|f_{all}|^s+|t_{all}|^s}$, and $\frac{|f_{fake|1}|^s}{|f_{all}|^s+|t_{all}|^s}$ where $|f_{all}|^s$ is the number of tweets, or shares, contributed by f_{all}.

Results are shown in Figure 1. For the *all* condition, based on (f,t), fake news could amount to be more than 40% of total news shares or as low as less than 3%. However, if we discard all domains with mixed factualness, prevalence drops substantially to between 1.3% and 20.1%. This suggests domains that are low in quality but not necessarily fake contribute to a large fraction of total articles shared. Further, we also observe that the fraction of fake news are comparable for the conditions *all-except-mixed* and *fake* except for ZDR, suggesting domains in these lists that are neither in the *fake* subcategory nor have mixed factualness contribute to very limited Twitter shares [4].

This analysis helps explain the divergent findings in the literature. While some studies raise significant concerns about the prevalence of fake news (Silverman, 2017), others claimed limited prevalence (Allcott, Gentzkow, and Yu, 2018). Here we see that drastically divergent conclusions can be reached even with the same Twitter data as a function of the fake and traditional news lists and fake-ness definitions (e.g. fake, mixed) one chooses to use.

5.2 Time Series Analysis

In this section, we first construct a time-series representing the fraction of fake news per day for each (f,t) from 3 different time periods (primary, general election, and after election) accounting for only Clinton tweets, only Trump tweets, and all tweets (for both candidates). We then compare these time-series from 3 distinct dimensions: i) correlation, ii) trend, and iii) effects of external events. Specifically, for each time interval i where $i \in \{primary, general\,election, after\,election\}$, given a groundtruth pair (f,t) and candidate c (where $c \in \{clinton, trump, both\}$), we write $|f|^s_{0,c}$, and $|t|^s_{0,c}$ as the total number of tweets, or shares, that mention c and contain urls from f and r at day 0 [5]. We then derive the time-series $P^i(f,t,c) = \{\frac{|f|^s_{0,c}}{|f|^s_{0,c}+|t|^s_{0,c}}, \frac{|f|^s_{1,c}}{|f|^s_{1,c}+|t|^s_{1,c}} \ldots\}$.

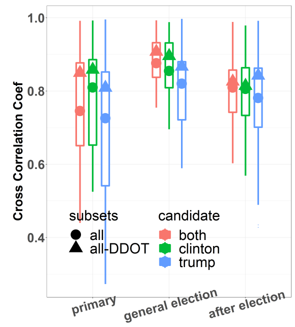
Time-series Correlation: We calculate correlation separately for each time period and candidate. For each c and i, given 2 groundtruth pairs $(f1,t1)$ and $(f2,t2)$ where $f_1 \neq f_2$ or $t_1 \neq t_2$, we compute the maximum normalized cross correlation coefficient and the corresponding time lag (Haugh, 1976) of $P^i(f_1,t_1,c)$ and $P^i(f_2,t_2,c)$.

We observe that the highest correlation scores of all pairwise comparisons occur at 0 lag, indicating that no single time-series is "ahead" or "behind" others. Correlation scores are plotted in Figure 2a. Normalized coefficients have a range between $\{-1,1\}$. As shown, correlation for $P(f_1,t_1,c)$ and $P(f_2,t_2,c)$ is the highest when $f_1 \equiv f_2$ but $t_1 \neq t_2$, indicating traditional news list selection (choosing ALEXA, MBFC(T), or VARGO) has little impact here. Further, we also note that certain fake news lists have considerably high correlation (e.g. ZDR and MBFC have correlation consistently higher than 0.9). Yet, DDOT diverges significantly from others (e.g. correlation between DDOT and POLIT is as low as 0.46 even with the same traditional news list).

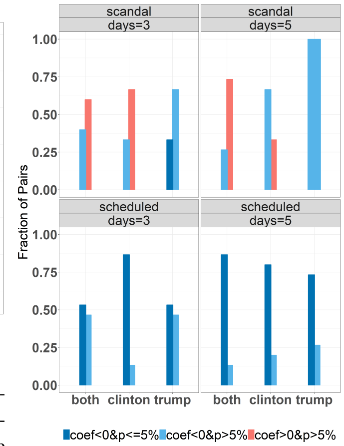
We further observe that the correlation is highest for the *general election* season (median correlation between the pairs for each candidate c are all above 0.8). Most efforts in fake news detection were motivated by the spread of fake news during the 2016 Election. This provides one potential explanation—fact-checkers and scholars could have had a stronger emphasis on the publishers that were active in this time frame, resulting in higher agreement.

[4]For robustness checks, we also measured fake news prevalence as the fraction of unique Twitter accounts that have shared at least 1 fake news article and observed a similar pattern.

[5]Here, we pick 2015-12-01, 2016-06-15, and 2016-11-09 as day 0 for primary, general election, and after election; and 2016-06-21, 2016-11-15, and 2017-01-01 as the last day.



(a) Cross correlation coefficient between pairs of lists for each time period. Circle and triangle points are the median coef scores when including or excluding DDOT.



(b) Level effect of scandals and scheduled events.

Table 2: Fraction of Fake News Per Day Time-series Trend Using Different GroundTruth. Column *majority trend* denotes the trend observed by a majority of pairs, column *majority frac* is the fraction of pairs in majority.

period	candidate	majority trend	majority frac	median β_1	least.congruent
primary	trump	positive	66.7%	0.04%	DDOT, POLTI
	clinton	stationary	60%	NA	
	both	stationary, positive	46.7%	NA, 0.02%	
general election	trump	positive	**100%**	0.03%	NA
	clinton	positive	**100%**	0.09%	
	both	positive	**100%**	0.07%	
after election	trump	negative	66.7%	-0.01%	AGZ, POLTI
	clinton	positive	80%	0.07%	
	both	positive	66.7%	0.05%	

Trend: Similar to prior work (Lazer et al., 2018), we are also interested in assessing whether there was an increase in fake news and to what degree findings would depend on the choice of groundtruth pairs. Here, we first use *statsmodel.seasonal_decompose* (Beveridge and Nelson, 1981) to deconstruct each time-series $P^i(f,t,c)$ into its components *trend*, *seasonal*, and *residual*. This is to remove the seasonality and residuals from the original time-series.

Next, we apply both Augmented Dickey-Fuller (ADF) and Kwiatkowski Phillips Schmidt Shin (KPSS), 2 commonly used methods to test for stationarity (Charemza and Syczewska, 1998), on the *trend* component of $P^i(f,t,c)$. If any one of the tests show that unit root is non-stationary, we run the linear regression model $y^i(f,t,c) = \beta_0 + \beta_1 * T + \varepsilon$ (where $y^i(f,t,c)$ contains the values from *trend*, and T is the time elapsed since the start of the time-series). Here, a positive β_1 suggests a rise of fake news. Finally, we assess whether trend analysis results for each candidate c and time period i using all pairs (f,t) are consistent.

Results are on Table 2. Column "majority trend" shows the trend result shared by the largest fraction of groundtruth pairs and column "majority frac" is the size of that fraction. Additional, median β_1 scores indicate the average fractional increase of fake news per day. As shown, conclusions for the *general election* are remarkably consistent: all lists pairs indicate an increase in fake news. Results for other time intervals are less congruent and we observe that DDOT and POLIT disagree the most with other lists (measured by the number of times a list diverges from "majority vote") in the *primary*, whereas AGZ and POLIT in *after election*.

Effects of External Events: Many prior studies examined media coverage of i) unexpected political events such as scandals (Puglisi and Snyder Jr, 2011) as well as ii) scheduled high-profile events such as the presidential debates (Scheufele, Kim, and Brossard, 2007). Such events are shown to have important effects on campaign news coverage. Here, we examine whether these 2 distinct categories of events have a temporary effect on the prevalence of fake news, specifically in the *general election* period.

We first obtain a list of scandals and planned key events of Trump, Clinton, or both that occurred in the general election from *ABC News* and *The Guardian*. The list, ordered chronically, includes: Republican nomination (07/18), Demoncrat nomination (07/28), Clinton "deplorable" and "pneumonia" scandals (09/09), first debate (09/26), Clinton email involving Wikileaks and Trump Hollywood tape scandals (10/07), second debate (10/09), Clinton email scandals involving the FBI (10/28, 11/06), and finally election day (11/08). Here, nominations, debates, and election day are assigned to *scheduled* and others to *scandal*.

Next, we use the autoregressive integrated moving average (ARIMA) time-series model (Stock, Watson, and others, 2003) to run interrupted time-series analysis and identify whether *scandals* and *scheduled* events are associated with level changes in the fraction of fake news per day for x days where $x \in \{3,5,7\}$. In our paper, we use *auto.arima*, a common ARIMA model selection function (Makridakis, Wheelwright, and Hyndman, 2008) from R's forecast library. Given a time-series, $P^i(f,t,c)$, and a set of external regressors (i.e. events), *auto.arima* selects the best ARIMA model based on the corrected Akaike information criterion (AIC). Here, we have 2 external regressors for each c. We denote $xreg^1_{c,T} = \{0,0,...1,1,1...\}$ where $xreg^1_{c,t} = 1$ if day t is within x days of the nearest *scandal* (after it has occurred) involving c. Similarly, we write $xreg^2_{c,T}$ for *scheduled* [6].

[6] If c is *both*, we only use events that involve both candidates.

A positive coefficient returned by *auto.arima* for $xreg^1_{c,T}$ would mean that *scandals* temporarily increase the total fraction of fake news per day. As shown in Figure 2b [7], regardless of the groundtruth selection, *scheduled* events generally contribute to a reduction of fake news. Though this does not mean planned events reduced the absolute volume of fake news, one possible explanation is mainstream media simply covered scheduled events much more, thus $\frac{|f|^s}{|f|^s+|t|^s}$ is smaller. Results for *scandal* are, however, more varied, suggesting that groundtruth pair selection has an impact on perceived effects of scandals.

5.3 Agenda-setting Priorities

In this section, we first use an iterative topic modeling process to extract issues, or topics, being covered by both fake and traditional news sites and assign each news article to its corresponding topic. then, we examine whether the choice of groundtruth pairs impacts agenda-setting conclusions.

Topic Modeling of News Articles Using Guided LDA: We use Guided LDA for topic modeling. It is an extension of the base LDA that allows sets of keywords to guide document topic assignment by increasing their "confidence" or weights (Jagarlamudi, Daumé III, and Udupa, 2012).

First, we use base LDA and manual labeling to extract seed words from news articles [8]. More specifically, we use *gensim* (Rehurek and Sojka, 2011) to generate several base LDA models [9]. We then select the model which has the optimal coherence score [10]. From it, we obtain the top 30 most representative words for each topic. Next, we manually inspect words and categorize them into coherent sets (i.e. topics). Using this approach, we obtain 409 unique seed words divided into 33 different sets. Next, we run the guided LDA implemented by Singh (2018) using the derived seed word sets, adjusting model's seed confidence to 0.25 and setting the number of total topics to 125 [11]. We filter out the subset of topics that lacked coherent themes and collapse topics that share the same human-interpretable theme into a single topic. This process results in 19 distinct topics. Finally, we assign each document into a single topic according to the maximum probability of its topic distribution. This topic is also later referred to as the document's *predicted* topic label.

[7] Trend results for when $x = 7$ is omitted due to space.

[8] We remove stop words in addition to lemmatizing and stemming non-stop words. Finally, we remove all articles that have fewer than 100 or more than 800 word tokens.

[9] The number of topics are $\{50,75,100,125,150\}$ respectively for the models. In addition, we set all models to ignore words that have a frequency of less than 100 or occur in more than 50% of total documents, and to generate bigrams for words that co-occurred at least 100 times or more.

[10] Coherence score for a topic is defined as the average of the pairwise word-similarity scores of the words in the topic (e.g. PMI) (Newman et al., 2010). A model's coherence score is calculated as the sum over its topic coherence scores.

[11] Here, we use perplexity score (Misra, Cappé, and Yvon, 2008), defined as log-likelihood per word, to determine the optimal number of topics given that *gensim* does not support coherence calculation for guided LDA.

Table 3: List of Topics, Fraction of Total Documents Accounted for, Most Weighted Keywords, and F1 Scores

topic	doc frac	most weighted tokens	f1
abortion	0.96%	woman abort life plan_parenthood issu punish femal	0.87
benghazi	0.60%	attack benghazi libya committe report secretari secur	0.75
c-health	0.86%	medic doctor releas report mental suffer pneumonia	0.75
climate	1.40%	climat coal environment industri land administr regul	0.89
wst	0.30%	speech wall_street talk ask issu transcript releas	0.82
d&i	0.75%	commun lgbt issu equal woman discrimin anti marriag	0.78
economy	4.4%	trade job china deal compani manufactur econom	0.79
election	20.3%	sander berni primari voter percent poll voter cruz	0.77
email	5.76%	email depart investig server classifi comey secretari	0.84
border	2.28%	immigr border mexico wall illeg deport mexican build	0.85
mid-east	3.86%	muslim islam israel isi terror terrorist attack unit syria	0.76
religion	1.14%	christian evangel church faith religi leader pastor pope	0.78
russia	1.81%	russia russian putin intellig hack offici govern	0.76
security	1.70%	iran china nuclear polici foreign deal nato secur	0.78
sexual	1.93%	woman accus alleg rape husband sexual claim sexual_assault	0.82

Topic Modeling Quality Assessment and Selection: For each topic, we randomly sample 0.2% of its total documents (or 10 if the size of a topic is small). This gives us 434 unique documents. We also sample 0.2% documents from the articles not included in the 19 topics. This results in 525 documents. Finally, we shuffle and publish the 1K $(434+525)$ documents on MTurk for crowdsourced labeling [12].

We assign 3 independent workers to categorize each document [13]. For each article, we mark its *manual* topic according to the majority vote (e.g. 2 out of 3 workers selected the same primary category) [14]. Next, for each topic, we calculate its precision, recall, and f1 scores using the *manual* and *predicted* topic labels. We filter out the topics that have an f1 score of < 0.75. This process produces 15 distinct topics which account for 49% of total news articles. Table 3 contains a summary on the list of topics ordered alphabetically. For each topic, we provide its name, its prevalence, most weighted keywords, and F1 score. As shown, *election* is the most prevalent topic accounting for 20.3% of total news articles, followed by Clinton's email scandal, and the economy.

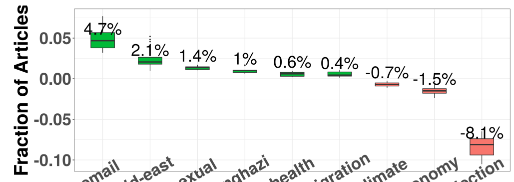
Agenda-setting Priorities: Next, we assess whether groundtruth choice affects the perceived agenda-setting difference between fake and mainstream news.

[12]The success of a crowdsourcing task relies heavily on the right mechanisms to ensure worker qualifications. We require that workers: 1) reside in the U.S. 2) have successfully completed at least 1,000 HITs; and 3) have an approval rate of at least 98%.

[13]Workers are given a list of categories (19 topics listed in Table 3 + 1 *none of the above* option) to choose from and are instructed to identify and select a single *primary* category of a given article. They are also instructed to choose *none of the above* option if the primary topic of an article is unclear. We use Krippendorff's alpha (Hayes and Krippendorff, 2007) to measure intercoder reliability. It is 0.62, which means a moderate agreement.

[14]Articles that do not have a majority is labeled as *unknown*. We observe 46, or 8.6% *unknown* documents. Note, *unknown* documents differ from *none of the above*.

Figure 3: Relative agenda priority difference between fake and traditional news. Y-axis is fraction of fake news articles on topic i subtract by the fraction of traditional news articles on i. Topics colored in green indicate a higher priority by fake news.



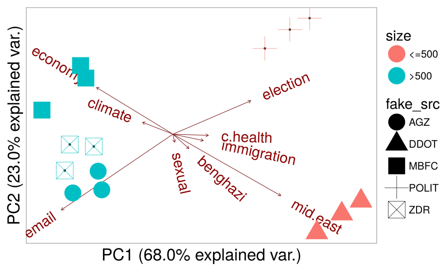
For a groundtruth pair (f,t), we derive the following topic distributions $K_{(f,t)} = \{k^1_{(f,t)}, k^2_{(f,t)} ... k^{16}_{(f,t)}\}$ and $L_{(f,t)} = \{l^1_{(f,t)}, l^2_{(f,t)} ... l^{16}_{(f,t)}\}$ where $k^i_{(f,t)}$ and $l^i_{(f,t)}$ are the fractions of fake and traditional news articles on topic i respectively, and $\sum_{i=1}^{16} k^i_{(f,t)} = 1$, $\sum_{i=1}^{16} l^i_{(f,t)} = 1$.

Then, for each topic i in $\mathscr{I}$ (where $\mathscr{I}$ is the entire set of topics), and all groundtruth pairs (F,T), we apply Student's T-test on $K^i(F,T)$ and $L^i(F,T)$ to determine whether the difference in mean is statistically significant between these 2 distributions (here, $K^i(F,T) = \{K^i(f1,t1), K^i(f2,t1) ... K^i(f5,t3)\}$). In other words, we assess whether fake news sites have published significantly more or fewer articles (measured using normalized fractions) on certain topics than traditional news sites and vice versa. We observe a significant difference in 9 topics. For instance, the average fraction of traditional news articles focusing on *election* is 22.5%, while the average is less than 15% for fake news articles. Traditional news sites are also more concentrated on topics including *economy* and *climate*. Fake news sites, on the other hand, spend a considerable fraction, approximately 10%, of all articles on Clinton's *email* scandal alone, twice that of traditional news sites. Fake news sites also place a stronger emphasis on topics such as *sexual* scandals (mostly related to Bill Clinton), and Hillary's pneumonia and claims of early onset dementia.

For each pair (f,t), we calculate the difference distribution $D_{(f,t)} = \{d^1_{(f,t)}, d^2_{(f,t)} ... d^9_{(f,t)}\}$ where $d^1_{(f,t)} = k^1_{(f,t)} - l^1_{(f,t)}$. We then plot $D^{\mathscr{I}}_{(F,T)}$ in Figure 3. Notably, the data points of $D^{\mathscr{I}}_{(f,t)}$ consistently stay above or below the horizontal $y = 0$ line. In other words, the assessment as to whether a topic had been more central to the coverage of fake news outlets compared to that of traditional news news sites is robust to the choice of groundtruth labels.

Groundtruth Difference Using Factor Analysis: Here, we provide a more in-depth analysis of how topics contribute to the variance in reporting by fake and traditional news sites of different (f,t) pairs. We apply Principal Component Analysis (Wold, Esbensen, and Geladi, 1987) to $D^{\mathscr{I}}_{F,T}$ and extract the first 2 principal components (the first and second component explains 68% and 23% of the total variance). The resulting biplot is shown in Figure 4. We see that `MBFC`, `ZDR`, and `AGZ` are more similar in their topic distributions. In comparison, fake sites in `POLIT` have a higher

Figure 4: PCA plot for topic fractional difference distribution between fake and traditional news described in Section. 5.3 Fake news lists are marked by shape.



fraction of articles on *election*. One possible explanation is that this list is specifically created to reduce election-related fake news (Politifact staff, 2018). Additionally, we also see that fake new sites in DDOT have a higher priority for scandals and controversial issues including *benghazi* and *sexual*, perhaps due to *Daily Dot* being a social news site focused on fake news sites that wrote more entertaining false content.

6 Conclusion

In this paper, we first provided a comprehensive overview of the publicly available lists of fake and mainstream news sites. We showed that these lists have divergent labeling processes and very few domains in common. In addition, we illustrated that the perceived prevalence of fake news varies substantially based on groundtruth choice. Further, we also observed an increase of fake news in the general election season regardless of the groundtruth selection, and that scheduled events contributed to a temporary reduction of fake news (conclusions for scandals were more mixed). Finally, after an iterative topic modeling process with considerable manual efforts from the researchers, we demonstrated that agenda-setting priority differences between fake and mainstream news sites remain relatively robust to the groundtruth difference. Overall, our results suggest groundtruth selection have moderate to limited (or explainable) impact on downstream analysis in i) prevalence, ii) temporal characteristics, and iii) agenda-setting priorities.

There are several caveats to our study. First, our analysis of groundtruth difference and its impact is limited to domain-level labels. There are more granular datasets that annotate at article even sentence-level. Second, understanding how groundtruth difference affect the performance of automated fake news classifiers may also provide valuable insights. Lastly, we would like to emphasize that groundtruth labels are indeed changing through time, future work should include methods that track and evaluate these changes.

Guidance on Groundtruth selection: Where do we go from here? How can we make progress as a research community despite the lack of agreement between fake news lists as to which domains should be considered fake? Our findings can be leveraged to provide guidance. First, researchers need to consider whether an analysis is *directly* affected by list size, as in the case of prevalence. Other types of analysis that depend on the nature of the fake news domain (as opposed to counts) are a lot more robust to the choice (e.g. temporal and topical analysis).

Table 4: A domain's likelihood of being listed by a source given its i) ideology, ii) subcategory, iii) age, and iv) popularity.

Independent Variables	(DDOT)	(POLIT)	(AGZ)	(MBFC)	(ZDR)
ideology_conservative	-0.007	0.049	0.125^{*}		0.006
ideology_unknown	-0.006	0.027	0.170^{**}		0.126
subtype_bias	-0.664^{***}	-0.480^{***}	-0.375^{***}	0.008	
subtype_clickbait	-0.677^{***}	-0.462^{***}	-0.373^{***}	-0.128^{*}	
subtype_conspiracy	-0.686^{***}	-0.442^{***}	-0.370^{***}	0.021	
subtype_hate	-0.703^{***}	-0.452^{***}	-0.348^{***}	-0.251^{***}	
subtype_junksci	-0.705^{***}	-0.460^{***}	-0.388^{***}	0.055	
subtype_rumor	-0.704^{***}	-0.420^{***}	-0.408^{***}	-0.148	
subtype_satire	-0.704^{***}	-0.345^{***}	-0.284^{***}	0.047	
subtype_unknown	-0.703^{***}	-0.412^{***}	-0.268^{***}	0.295^{***}	
subtype_unreliable	-0.703^{***}	-0.512^{***}	-0.511^{***}	-0.178^{***}	
popularity	-0.0004	-0.002	0.023^{***}	0.048^{***}	0.042^{***}
age_in_year	-0.0001	-0.009^{***}	-0.024^{***}	0.008^{***}	-0.003
Observations	1,645	1,645	1,645	1,645	1,645
p-value	0.62	0.21	0.24	0.18	0.17

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

The second consideration relates to which lists one should use for evaluation. We recommend MBFC and POLIT for analysis related to the 2016 Election. To elaborate, we first observe that the choice of traditional news lists seems to not matter–reducing the effort to carry out research. Second, we also see consistent clustering of lists. That is, MBFC, AGZ, and ZDR are commonly clustered together (clustering of lists in the topic analysis latent space). Within the 3 lists, MBFC marks the upper bound on prevalence analysis and has a substantive labeling process–making it the best choice of the 3. Next, POLIT and DDOT are rather distinct from the rest. As previously stated, POLIT was created independently to study and combat the most popular fake news in the 2016 Election, therefore it is likely a good evaluation list for studies concerning the Election. DDOT, however, heavily referenced existing lists and provided little justification for the subset of domains they handpicked. Finally, we believe expanding as a function of list clustering, i.e. considering the next most distinct list, can help explore this data space and we hope that our meta-analysis (e.g. annotation and quality measures) provide further guidance here.

7 Appendix

Regression Model for Domain Inclusion For a given domain i that's listed by at least one f where $f \in \{ZDR, MBFC, AGZ, DDOT, POLIT\}$, let the binary variable $y_{i,s} = \{0,1\}$ denote whether domain i exists in the list of fake news sites f. We fit regression model for each f using *ideology*, *subcategory*, *popularity*, and *age* as the explanatory variables.

$$y_f = \beta_0 + \beta_1 ideology + \beta_2 subtype + \beta_3 popularity + \beta_4 age + \varepsilon_i$$

Results are summarized on Table 4.

Regression Model for the Time of Addition We first use web.archive.org and authors' websites to obtain 3 timestamped snapshots [15] of lists ZDR, MBFC, and DDOT. Let i be a website that was added to ZDR in one of its 3 snapshots and remained on the list thereafter, we determine i's preferred *ideology*, *subcategory*, *popularity*, and *age*. Let the variable $y_{i,zdr} = \{0,1,2\}$ denote whether domain i was added in the 1st, 2nd, or 3rd version of ZDR, we fit the following regression model:

$$y_{zdr} = \beta_0 + \beta_1 ideology + \beta_2 subtype + \beta_3 popularity + \beta_4 age + \varepsilon_i$$

We repeat the same procedure for DDOT and MBFC. Regression results table is omitted due to space.

Besides the addition of domains through time, we also looked into i) domain removals and ii) domains with changed subcategories. We observe very few to no removals [16]; same for changes of subcategories.

References

Alimonti, K. R., and Veridiana. 2018. fake news offers latin american consolidated powers an opportunity to censor opponents.

Allcott, H., and Gentzkow, M. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31(2):211–236.

Allcott, H.; Gentzkow, M.; and Yu, C. 2018. Trends in the diffusion of misinformation on social media. *arXiv preprint arXiv:1809.05901*.

Balmas, M. 2014. When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. *Communication Research* 41(3):430–454.

Beveridge, S., and Nelson, C. R. 1981. A new approach to decomposition of economic time series into permanent and transitory components. *Journal of Monetary economics* 7(2):151–174.

Bode, L.; Budak, C.; Ladd, J. M.; Newport, F.; Pasek, J.; Singh, L. O.; Soroka, S. N.; and Traugott, M. W. Forthcoming. *Words That Matter: How the News and Social Media Shaped the 2016 Presidential Campaign*. Washington, D.C.: Brookings Institution Press.

Bovet, A., and Makse, H. A. 2019. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications* 10(1):7.

Budak, C. 2019. What happened? the spread of fake news publisher content during the 2016 u.s. presidential election. WWW '19.

Charemza, W. W., and Syczewska, E. M. 1998. Joint application of the dickey-fuller and kpss tests. *Economics Letters* 61(1):17–21.

Croft, M., and Moore, R. 2017. Checking what students know about checking the news.

[15] December 2016, June 2017, and December 2017 with each separated by 6 months

[16] The exception being DDOT: in late 2016, DDOT contained 98 websites; it then removed a substantial number of sites and reduced its size to 25 in mid-2017; its latest version has a size of 175. No explanation was given for each change.

Darwish, K.; Magdy, W.; and Zanouda, T. 2017. Trump vs. hillary: What went viral during the 2016 us presidential election. In *International Conference on Social Informatics*.

Doshi, A. R.; Raghavan, S.; Weiss, R.; and Petitt, E. 2018. The impact of the supply of fake news on consumer behavior during the 2016 us election.

Farmer, L. S. 2017. Dont get faked out by the news: Becoming an informed citizen. In *The Fifth European Conference on Information Literacy (ECIL)*, 174.

Farte, G.-I., and Obada, D.-R. 2018. Reactive public relations strategies for managing fake news in the online environment.

Figueira, Á., and Oliveira, L. 2017. The current state of fake news: challenges and opportunities. *Procedia Computer Science* 121:817–825.

Goodall, D. W. 1966. A new similarity index based on probability. *Biometrics* 882–907.

Granskogen, T. 2018. Automatic detection of fake news in social media using contextual information. Master's thesis, NTNU.

Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2018. Fake news on twitter during the 2016 us presidential election. Technical report.

Guess, A.; Nyhan, B.; and Reifler, J. 2018. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign. *European Research Council* 9.

Guo, L., and Vargo, C. 2018. "Fake News" and Emerging Online Media Ecosystem. *Communication Research* 009365021877717.

Haugh, L. D. 1976. Checking the independence of two covariance-stationary time series: a univariate residual cross-correlation approach. *Journal of ASA* 71.

Hayes, A. F., and Krippendorff, K. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures* 1(1):77–89.

Helmstetter, S., and Paulheim, H. 2018. Weakly supervised learning for fake news detection on twitter. 274–277. IEEE.

Horne, B. D., and Adali, S. 2017. This Just In. *arXiv:1703.09398 [cs]*. arXiv: 1703.09398.

Horne, B. D.; Dron, W.; Khedr, S.; and Adali, S. 2018. Assessing the News Landscape. 235–238. Lyon, France: ACM Press.

Jagarlamudi, J.; Daumé III, H.; and Udupa, R. 2012. Incorporating lexical priors into topic models. 204–213. ACL.

Koulolias, V.; Jonathan, G. M.; Fernandez, M.; and Sotirchos, D. 2018. Combating misinformation: An ecosystem in co-creation.

Kucharski, A. 2016. Post-truth: Study epidemiology of fake news. *Nature* 540(7634):525.

Lazer, D. M.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; et al. 2018. The science of fake news. *Science* 359(6380):1094–1096.

Leetaru, K., and Schrodt, P. A. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, 1–49. Citeseer.

Lima, L.; Reis, J.; Melo, P.; Murai, F.; Araújo, L.; Vikatos, P.; and Benevenuto, F. 2018. Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. *arXiv:1807.03688*.

Main, T. J. 2018. *The Rise of the Alt-Right*. Brookings Institution Press.

Makridakis, S.; Wheelwright, S.; and Hyndman, R. 2008. *Forecasting methods and applications*. John wiley & sons.

Marcon, A. R.; Murdoch, B.; and Caulfield, T. 2017. Fake news portrayals of stem cells and stem cell research. *Regenerative medicine* 765–775.

McCombs, M. E., and Shaw, D. L. 1972. The agenda-setting function of mass media. *Public opinion quarterly* 36(2):176–187.

Misra, H.; Cappé, O.; and Yvon, F. 2008. Using lda to detect semantically incoherent documents. 41–48. Association for Computational Linguistics.

Mitra, T., and Gilbert, E. 2015. CREDBANK: A Large-Scale Social Media Corpus with Associated Credibility Annotations. 10.

Mukerji, N. S. 2018. A conceptual analysis of fake news.

Nelimarkka, M.; Laaksonen, S.-M.; and Semaan, B. 2018. Social media is polarized, social media is polarized: Towards a new design agenda for mitigating polarization. 957–970. ACM.

Newman, D.; Lau, J. H.; Grieser, K.; and Baldwin, T. 2010. Automatic evaluation of topic coherence. 100–108. Association for Computational Linguistics.

Palen, L., and Hughes, A. L. 2018. Social media in disaster communication. In *Handbook of Disaster Research*. Springer. 497–518.

Pennycook, G.; Cannon, T.; and Rand, D. G. 2018. Prior exposure increases perceived accuracy of fake news.

Politifact staff. 2018. Politifact guide to fake news websites and what they peddle.

Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; and Stein, B. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Poynter Institute. 2019. International Fact Checking Network. [Online; accessed -04-November-2019].

Puglisi, R., and Snyder Jr, J. M. 2011. Newspaper coverage of political scandals. *The Journal of Politics* 73(3):931–950.

Rapoza, K. 2017. Can fake news impact the stock market?

Rehurek, R., and Sojka, P. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Masaryk University* 3(2).

Rini, R. 2017. Fake news and partisan epistemology. *Kennedy Institute of Ethics Journal* 27(2):E–43.

Schaedel, S. 2018. Websites that post fake and satirical stories. factcheck.

Scheufele, D. A.; Kim, E.; and Brossard, D. 2007. My friend's enemy: How split-screen debate coverage influences evaluation of presidential debates. *Communication Research* 34(1):3–24.

Shao, C.; Ciampaglia, G. L.; Flammini, A.; and Menczer, F. 2016. Hoaxy. *WWW '16 Companion*. arXiv: 1603.01511.

Shao, C.; Ciampaglia, G. L.; Varol, O.; Flammini, A.; and Menczer, F. 2017. The spread of fake news by social bots. 96–104.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19(1):22–36.

Shu, K.; Wang, S.; and Liu, H. 2018. Understanding user profiles on social media for fake news detection. In *2018 IEEE Conference on (MIPR)*, 430–435. IEEE.

Silverman, C. 2016. Here are 50 of the biggest fake news hits on facebook from 2016.

Silverman, C. 2017. The fake news watchdog.

Starbird, K. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *ICWSM*, 230–239.

Stock, J. H.; Watson, M. W.; et al. 2003. *Introduction to econometrics*, volume 104. Addison Wesley Boston.

Tandoc Jr, E. C.; Lim, Z. W.; and Ling, R. 2018. Defining fake news a typology of scholarly definitions. *Digital Journalism* 6(2):137–153.

Van Zandt, Dave. 2018. Media bias/fact check (mbfc news) about.

Vargo, C. J.; Guo, L.; and Amazeen, M. A. 2018. The agenda-setting power of fake news. *new media & society* 20(5):2028–2049.

Vikash, S. 2018. Guidedlda.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science* 359(6380):1146–1151.

Wang, W. 2017. "Liar, Liar Pants on Fire". *arXiv:1705.00648*.

Wardle, C. 2017. Fake news. its complicated. *First Draft News*.

White, Nicholas. 2018. The daily dot — Wikipedia, the free encyclopedia. [Online; accessed 27-October-2018].

Wikipedia contributors. 2019. Alexa internet — Wikipedia, the free encyclopedia. [Online; accessed 4-May-2019].

Wineburg, S. 2016. Stanford researchers find students have trouble judging the credibility of information online.

Wold, S.; Esbensen, K.; and Geladi, P. 1987. Principal component analysis. *Chemom. Intell. Lab. Syst.* 2(1-3):37–52.

Zimdars, M. 2016. My fake news list went viral. but made-up stories are only part of the problem. *The Washington Post*.