A Comparative Evaluation of Methods for Discovering Potential Fake News Domains

Anas Elghafari
New York University
anas.elghafari@nyu.edu

Zhouhan Chen
New York University
zc1245@nyu.edu

Aécio Santos
New York University
aecio.santos@nyu.edu

Sonia Castelo
New York University
s.castelo@nyu.edu

Thais Almeida
Federal University of Amazonas
Manaus, Amazonas
tga@icomp.ufam.edu.br

Juliana Freire
New York University
juliana.freire@nyu.edu

ABSTRACT

Online fake news and political misinformation have attracted research attention in recent years owing to their influence on important events in several countries. Numerous research efforts have focused on developing methods for classifying fake news, both on social media and on the Web. These approaches have typically focused on identifying the predictive features of fake news and have usually been evaluated on relatively small sets of data. However, the scale of web content and the impossibility of an exhaustive application of any classification method to all sites raise the problem of how to discover the sites that are suitable input for the various classification methods. This discovery problem has been largely overlooked in the research literature on misinformation.

In this work, we develop and evaluate several automated methods for discovering web sites that are good candidates for further examination, i.e. potential fake news sites. The methods typically start from a set of seed domains and then leverage a variety of web APIs and clustering techniques to discover related domains. In applying the proposed methods to a set of 477 known fake news domains, we discover a total 20,937 sites, from which we crawl 1.8 million pages.

We conduct two experimental evaluations of the collected sites and pages, one using an SVM classifier of fake news web pages and one using manual sampling, to evaluate the relevance of each method's results, and compare the yield, characteristics, and pairwise similarity of the proposed methods. We show that, while there is some overlap in the results returned by the different methods, those result sets are largely heterogeneous, and a recall-maximizing strategy for discovering fake news sources will likely require an ensemble of such methods.

CCS CONCEPTS

• **Information systems** → *Retrieval models and ranking*; *Evaluation of retrieval results.*

CIKM'19, 2019

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.

KEYWORDS

Misinformation, Fake News, Online Search, Website Discovery, Domain Discovery, Fake News Discovery, Online News

ACM Reference Format:
Anas Elghafari, Zhouhan Chen, Aécio Santos, Sonia Castelo, Thais Almeida, and Juliana Freire. 2021. A Comparative Evaluation of Methods for Discovering Potential Fake News Domains. In *Proceedings of International Conference on Information and Knowledge Management (CIKM'19).* ACM, New York, NY, USA, 15 pages.

1 INTRODUCTION

Fake news and misinformation have played an important role in shaping public opinion and driving events in several countries [1]. This has been enabled by the web infrastructure that makes it easy and cheap to create web sites and publish content, and by the increasing role of social media platforms which give content from otherwise obscure sites the chance to propagate to audiences far beyond what those sites could organically reach on their own.

While misinformation has attracted substantial attention in the Computer Science and Data Science research communities, the problem is still not well understood [14]. Moreover, even human agents can have difficulty in discerning fake from non-fake news [8, 9]. Thus, the challenges for fact-checking efforts are both qualitative and quantitative: Besides the difficulty inherent in parsing and cross-referencing conflicting sources and claims, the ease of publishing content on the Web has led to orders of magnitude increase in the number of news sources and volume of content. Automated methods that identify potential misinformation sources can aid manual fact-checking by providing contextual information and limiting the volume of content that the human fact-checkers need to consider.

Most of the research efforts in this area have focused on developing classification approaches that can help identify misinformation. We briefly review some of those approaches in the Related Work section. However, when one considers the question of how to deploy such classifiers to find the fake news sources in the real world, a problem quickly becomes apparent: Where to start from? That is, given the scale of the web, some selection strategy is needed to identify the content most worthy of the limited evaluation resources.

This problem, how to discover content relevant for a given question, has received some attention in the context of web crawling research [3, 21, 24]. However, outside of one work we know of that focused on discovering content for fact checking [30], the discovery problem has been largely overlooked in misinformation research in favor of more focus on the classification problem. We believe that, given the impracticality of exhaustive evaluation of web content, real-world deployment of any classification system will require a robust discovery component.

In this paper, we focus on the domain discovery problem as independent of, and upstream from, the classification problem. The research questions this work investigates are:

- What web resources and APIs can be used to develop domain discovery methods?
- How do the discovery results from the various methods compare with each other?
- What is the quality of the various discovery methods' results (in terms of potential fake news domains discovered).

The main contributions of this paper are the following:

- We propose nine discovery methods that use a variety of APIs (Twitter's, Google's, Bing's). The methods are grouped into three classes based on input type: Domain based, content based, and social network based.
- We compile a list of 477 Fake News sites from an array of fact-checking and journalistic organizations. We apply the proposed discovery methods to these sites, yielding discovery result sets of 20,937 previously unknown domains
- We conduct a comparative evaluation of the result sets, and we evaluate against the set of known Fake News domains and a set of mainstream news domains. For the vast majority of discovered domains for which we do not have the ground truth, we carry out a classifier-aided evaluation as well as manual labeling of samples from the discovery sets. We report the results for each individual method and each class of methods.

2 RELATED WORK

In this section, we briefly review some of the previous work on the problem of classifying fake news on the Web and the approaches used for content discovery.

2.1 Fake News Classification

While a lot of the attention of the community has focused on detecting fake news on social media, several research works in recent years have took on the problem of detecting web-based fake news. For instance, Horne et al. [12] used the BuzzFeed 2016 dataset [28] and extracted features based on readability scores, complexity of the sentence structure, part of speech of the words used, etc. They found that fake news tends to be simpler and more repetitive. Their features were very effective for the task of differentiating satire from real news (0.91 accuracy), but less so for differentiating fake news from real (0.78 accuracy). Perez-Rosas et al. [19] also considered writing style and proposed the use of linguistic features that capture both content-based and content-agnostic aspects of web pages. These included n-grams, punctuation and psycho-linguistic features which capture deception, readability (e.g. proportion of long words), and syntax. They experimentally showed that their model attains accuracy of up to 0.76 on two publicly-available datasets.

Other approaches explored both network aspects of the web pages and user engagements. Fairbanks et al. [10] investigated whether credibility and bias can be assessed using content-based and structure-based methods. The structure-based method constructs a reputation graph where each node represents a site, and the edges represent mutually linked sites, as well as shared CSS, JavaScript, and image files.

Shu et al. [27] proposed the TriFN framework, which uses features about the news sites' partisan biases, linguistic features of news articles, and the credibility of the twitter users who are engaging with the content (building upon work by Castillo et al. [6]). TriFN combines latent matrix representations with a semi-supervised approach to make predictions. They found that content-related features are not effective on their own, but when combined with features about the site (partisan bias) and the users' credibility, they can attain high accuracy.

There have also been approaches that attempt to distinguish between different types of fake news. Rashkin et al. [25] compared the language of real news with that of satire, hoaxes, and propaganda. They then explored the feasibility of predicting the reliability of the news article into four categories: trusted, satire, hoax, or propaganda. Their Max-Entropy classifier with L2 regularization on n-gram tf-idf feature vectors, applied on pages from eight distinct sites, resulted in F1 scores of 0.65. Hosseinimotlagh and Papalexakis [13] proposed a tensor-based model that captures latent relations between articles and terms, as well as spatial relations between terms, to cluster fake news into different categories. However, they did not attempt to distinguish between fake and real news, focusing instead on categorizing fake news into different classes.

Other works also examined the potential of linguistic features to predict content quality. Louis and Nenkova [16] developed a model to predict the quality of news articles about science based on features such as sentence structure. Pitler and Nenkova [22] considered features that provide a good indication of readability.

2.2 Domain Discovery on the Web

Several techniques have been proposed to discover domain-specific web content. These include 1) *search-based discovery* techniques which rely on search engine APIs to find web pages based on query terms or based on similarity to other web pages [24, 29, 30] and 2) *crawling-based discovery* techniques, which use the Web link structure to explore new content by automatically downloading and recursively following links extracted from the discovered web pages [2, 3, 18]. While crawling-based discovery is more scalable, search-based discovery methods have the advantage of enabling more targeted and efficient content discovery while requiring less computational resources than crawling approaches. In what follows, we mention some recent works that are either focused on discovery of fake news or can potentially be used for that.

Wang et al. [30] proposed a pipeline to discover documents that are relevant for fact-checking articles. They started from a set of fact checking articles and explored ways to find other articles that are relevant to the fact-checked claims. They then performed stance classification (deciding whether the discovered documents support or contradict the fact-checked claims). The discovery methods explored include querying search engine APIs for the title of the fact checking article, querying for a concatenation of the named entities in the article, as well as using Google's internal click graph data. They evaluated the recall of their methods against a gold-standard data produced by skilled workers performing open-ended searches. Their approach overlaps with ours in the use of search engine queries on titles of seeds, but it differs in its use of fact checking articles, rather than fake news sources, as seeds. As such, their approach is not designed to be a part of fake news detection system. Instead, the problem of fake news classification is sidestepped in favor of a *stance classification* problem where the task is determining if an article supports or contradicts a claim.

Beyond the specific problem of misinformation, the task of domain discovery for crawling purposes has received some attention in the literature. For example, Vieira et. al. [29] proposed a system that uses relevance feedback to gather seeds to bootstrap focused crawlers. It submits keyword search queries to a search engine API; extracts keywords from the result pages classified as relevant; and uses these keywords to construct new search queries. Disheng et. al. [24] presented a discovery pipeline that uses not only keyword search to find websites containing product specifications, but also leverages APIs to search for backlinks. A similar strategy that combines crawling and backlink search was also used by Barbosa et al. [3] to discover bilingual websites.

More recently, Pham et al. [21] proposed DISCO, an unified framework for discovering relevant websites given only a initial set of seed websites that belong to the domain of interest. Their approach is able to combine multiple discovery methods and can also learn which method works best for a given domain using a multi-armed bandit approach. While they evaluate their approach using multiple domains, they only experiment with search and crawling based discovery methods. This work is complementary to the work of Pham et al. [21] in the sense that we study the impact different discovery methods in the context of potential fake news sites discovery. In addition, the novel social-media based discovery methods proposed in the work could potentially be integrated into DISCO to improve its performance for domains that require timely content discovery. In earlier work, Pham et al. [20] has shown that new relevant content can be discovered by re-crawling previously crawled websites. Their work focuses on re-crawling strategies that maximize the yield of relevant web pages. This line of work could be incorporated in later stages of our pipeline: once potential fake news sites are discovered, they could then be monitored in order to timely detect fake news articles.

3 PROBLEM DEFINITION

Defining and Delimiting "Fake News"

We note that there is no widely-accepted definition for fake news. Here, we focus on all types of active political misinformation that goes beyond mainstream partisan bias, and consider as potential fake news domains not only sites that publish fabricated stories, but also sites that display a pattern of promoting misleading headlines, extremely-distorted content, and conspiracy theories. We view it as the task of fact checkers to draw the exact line between those different categories. From a fake news domain discovery perspective, all of these sites are worthy of further analysis, whether by automated classifiers or by human fact checkers.

It is important to note that our work focuses on discovering fake news domains, not fake news content. The need for this distinction becomes clear when one considers the problem of fake news that might occasionally be published by established, mainstream media organizations. For our purposes, if a news site is established and commonly accepted as a legitimate news source, then it is not what we are looking for, even if there are some instances of that site contributing to the spread of a fake news story. Our focus is on discovering the unknown news domains that are not being examined but ought to be. Hence, we adopt the somewhat orthogonal labels of "Fake News" and "Mainstream News".

Defining the Fake News Domain Discovery Task

Given a set of seed domains S containing Known Fake News Domains (KFNDs) and a domain discovery method M, *M-DD* refers to the set of discovered domains resulting from applying method M to each of the KFNDs and aggregating the results. Our evaluation of M centers around the number and proportion of Potential Fake News Domains in *M-DD*. We will also consider the number and proportion of Known Mainstream Domains that show up in the results, the number of political domains generally, and the cost -both in terms of money and API calls- of collecting those results.

4 COLLECTION OF DISCOVERY SEEDS

Known Fake News Domains

To construct the set of known fake news sites, we aggregated lists from Politifact [11] (taking only the sites labeled "fake news" and "imposter site"), Buzzfeed 2016 and 2017 sets [4], and Opensources.co[1] (taking only news labeled as "fake" and "fake news"). This aggregation yielded a list of 477 domains, referred as set S. However, as we will see below, the majority of those are no longer operational.

Content from the Known Fake News Domains

Several of the methods we explore here do not start from the KFNDs themselves, but rather start from articles from KFNDs. This is the case with all the methods where we interface with search engine APIs. This creates the need to obtain a sample of content from each KFND. Since we wanted to start from content that is popular, and thus more likely to be re-published by other sites, we decided to use Twitter to collect popular content from KFNDs. We queried Twitter's search API using the names of KFNDs, ranked the returned results by number of re-tweets, and took the top 20 articles (for many domains, not that many were found). In total, we found 123 KFNDs that are being shared on Twitter, and collected a total 1,428 articles from those domains.

[1] http://www.opensources.co

User Networks for Fake News Domains
We started from the known fake news domain S, and queried Twitter search API for tweets that contain at least one of the domains. This yielded 80,049 original tweets from 25,175 unique users. We then used Twitter Rest API to collect the past 200 tweets of each user. For each fake news site in S, we checked if it had been tweeted by at least 20 users. If so, we constructed an undirected, weighted graph $G < V_U, V_D, W >$, where V_U contains all user nodes and V_D contains all domain nodes. $W_{u,d}, u \in V_U, d \in V_D$ encodes all domain-user interaction, and weight $W_{u,d}$ = number of tweets user u sends that contain domain d. We chose threshold 20 because this number has worked well empirically in previous work on Twitter botnet detection [7]. We also removed edges that have weights smaller than or equal to three, as such edges usually represent random interactions instead of meaningful signals. We found that pruning the graph has little effect on the number of known fake news discovered, while at the same time greatly reduced the computational complexity. Table 1 shows the sizes of user networks constructed for the top known fake news domains.

Negative Instances - Known Mainstream News Domains:
While our focus is on discovery from known fake news domains, part of our evaluation (in 6.2) involved assessing what might be termed 'negative recall' for each method. That is, how many of known mainstream news domains showed up in the discovery set for each method. This requires a list of the most common mainstream news domains. For this we started from *Alexa's top 500 news sites* list[2] and manually excluded non-politics news domains (e.g. weather, sports, finance news domains). We refer to the resulting set as the Known Mainstream Domains (KMDs).

Seed domain	User node count	Domain node count	Edge count
worldtruth.tv	22	624	864
wnd.com	193	2346	7952
bipartisanreport.com	21	439	917
newspunch.com	53	969	2119
yournewswire.com	28	702	1011
infowars.com	239	2237	8186
beforeitsnews.com	62	1146	2183
breitbart.com	1618	6787	58816
politicususa.com	173	1355	5903
en-volve.com	29	452	972
naturalnews.com	298	4444	13026
realfarmacy.com	43	1381	2213
twitchy.com	181	1640	6238
thefreethoughtproject.com	88	1964	3924
clashdaily.com	53	818	1689
21stcenturywire.com	42	856	1803
conservativedailypost.com	42	709	1550
disclose.tv	80	1641	2993
dcclothesline.com	71	1033	2829
worldnewsdailyreport.com	36	922	1499
huzlers.com	23	398	618
activistpost.com	53	1309	2487
globalresearch.ca	96	1727	3736
lewrockwell.com	65	1454	3154
theblaze.com	283	2675	10572
neonnettle.com	119	2064	5430
redstate.com	126	1643	4550
veteranstoday.com	31	802	1259

Table 1: Statistics of 28 social network graphs: number of user node, number of domain node and number of edges. Each graph has one known fake news site that is connected to all user nodes. We call this node the seed domain.

5 DISCOVERY METHODS

We group discovery methods into three classes: (1) *domain-based* – those that take the domain as input seed, (2) *content-based* – those that use samples of content from the KFNDs as input, and (3) *network-based* – those that utilize user information to construct networks and find new sites using node similarity.

5.1 Domain-based Methods

Google Search Related Sites
Google has a functionality for related sites, which is accessible via the API. It takes as input a domain name and returns a list of domains that Google deems related to the input domain. We used this on each of the KFNDs, yielding a discovery set of 1,228 domains.

Alexa Overlap
Alexa Internet Analytics Company (part of Amazon) collects data from a sample of users that it uses to compile site-related data, including category, ranking, and relations to other sites in terms of backlinks or sample audience overlap. That data is available for a monthly fee. We queried the overlap data for each KFNDs, collecting the top 100 domains for each KFND. This yielded a total 3,586 domains.

Twitter Co-Sharing
Starting from each known fake news domain, S, we queried Twitter's API for tweets sharing links from S. This yielded a set of URLs from S as well as a set of users sharing those URLs. We used Twitter's API again to obtain most recent 1,000 or so tweets of each user, extract the domains of the links that occur in those tweets, and rank them by frequency of occurrence. We excluded the common social media and link shortening sites. From the remaining, We took the top 70 most frequent domains in that ranking as the domains discovered via S. As an example, Table 2 shows the domains discovered via *infowars.com* using this method (before filtering out the social media and link shortening domains)

5.2 Content-based Methods

Unlike the domain-based methods, the idea of content based methods is to leverage search engines to find sites publishing similar content to that published by the known fake news domains. As explained above, those methods require collecting samples of content (typically articles) from the seeds. The titles of the articles are then used as starting queries to discover candidate FN sites.

[2]https://www.alexa.com/topsites/category/News

Discovered Domain	Users Count	Links Count
infowars.com	33	4936
thegatewaypundit.com	27	322
dailycaller.com	25	280
breitbart.com	23	631
foxnews.com	23	168
zerohedge.com	22	273
fxn.ws	21	72
thehill.com	21	99
washingtonexaminer.com	20	83
truepundit.com	20	281
dailymail.co.uk	19	69
dailywire.com	18	150
ow.ly	17	62
washingtonpost.com	17	24
washingtontimes.com	17	67
hannity.com	15	52
trib.al	14	45
townhall.com	14	54
cnn.com	14	19
newswars.com	13	136
grrrgraphics.com	13	96
thefederalist.com	13	52
jwatch.us	13	42
cnbc.com	13	22
nytimes.com	13	28
latimes.com	12	13
theguardian.com	12	22
conservativetribune.com	12	175
politico.com	11	15
nationalreview.com	11	53
cbsnews.com	11	17

Table 2: Example of discovery with Twitter, starting from *infowars.com.* Top 30 domains among collected sample of sharers of infowars links, excluding common social media and link shorteners. Column 1 shows the number of infowars sharers in sample sharing the discovered domain, column 2 shows number of links shared

Google Custom Search API
We used Google's search API to query for the headlines of the collected articles. For each query we collected the top 20 returned results (Google's API returns the results 10 at a time and a request from the subsequent 10 count as an additional API query). We extracted the domain names from the set of collected search results, yielding a total of 2,004 domains

Bing Web Search API
We used Bing Search API to query for the headlines of the sampled content, collected the top 20 search results and extracting their domain names. The Bing search API allows us to submit the query both as boolean query and as an exact term search. We implemented and collected the results for both. Applied to all content collected from KFNDs, this exact term search yielded a discovery set of 6,641 domains, while the boolean queries yielded 12,886 domains.

Bing News API
Related to the Web Search endpoint, Bing offers an API endpoint for news search. The difference from the Web Search appears to be that returned results are limited to what Bing considers news sites. We evaluated this endpoint as a separate method. However, unsurprisingly, the numeric results from this methods appear to be very similar to the Bing Web Search with Boolean queries, with 0.94 Jaccard similarity coefficient between the two methods (Table 3)

5.3 Network-based Methods

The network-based methods leverage information from Twitter social networks. Using the social network graphs constructed in section 4, we attempted to identify domains that have similar node property to a known fake news domain. We implemented two different algorithms to measure node similarity: Jaccard's similarity and random walk. Those two methods are evaluated along with other discovery methods.

Jaccard's Similarity
To find nodes similar to our seed node, we define a function $score(d_1, d_2)$ to quantify the similarity between two domains. One commonly used similarity metric is Jaccard's similarity [15]. It measures the probability that both d_1 and d_2 have a feature i, for a randomly selected feature i that either d_1 or d_2 has. For any domain $d \in V_d$, we choose the feature to be the set of users connected to d, and write it as $\Gamma(d)$. Then our Jaccard's similarity is defined as:

$$score(d_1, d_2) = \frac{|\Gamma(d_1) \cap \Gamma(d_2)|}{|\Gamma(d_1) \cup \Gamma(d_2)|}$$

To select top k similar nodes, we sort scores in descending order and return top k domains. Jaccard's similarity only looks at directly-connected neighbors of a seed domain, therefore overlooking the multi-faced network structure.

Two-Step Random Walk on Bipartite Graph
A bipartite graph is a graph which contains two sets X and Y, and every edge must connect one element in X and one element in Y. Our original graph $G < V_U, V_D, W >$ is an example of a bipartite graph. Inspired by recent work in recommendation systems in identifying similar users based on common items [23, 26], we use the approach to identify candidate FN domains based on Twitter users shared with the known FN domains.

To walk on the graph, we construct a normalized transition matrix $T = D^{-1}W$, where D is the degree matrix. We initialize a probability vector p to represent distribution of resource. At step 0, $p_i = 1$, where i is the index of seed node, and $p_j = 0$, for any other index j. At step 1, $p = p \cdot T$, after which the resource is distributed from seed domain to Twitter users. At step 2, $p = p \cdot T$, after which the resource is collected from Twitter users to domains. To select top k nodes similar to the known fake news node, we extract k users corresponding to the highest values in vector p.

Random walk has a nice property that allows us to explore different areas of the graph, and to reach node more than one hop away from the seed node. Neighbor-based approach (such as Jaccard's similarity) cannot achieve this. We illustrate our random walk method in Figure 1.

6 EXPERIMENTAL EVALUATION

6.1 Pairwise Overlap of Discovery Results

The methods we have implemented vary greatly in input type, in number of domains yielded, and in cost. To better understand how they compare to each other, we measured the results overlap between each pair of discovery methods. For this we used Jaccard Similarity Coefficient, a similarity measure between two sets defined as ratio of the cardinality of the intersection to the cardinality of the union. Formally, given two discovery methods, M1, M2:

$$Jaccard(M1, M2) = \frac{|M1DD \cap M2DD|}{|M1DD \cup M2DD|} \tag{1}$$

where M*i*DD denotes the set of domains discovered through M*i*. Table 3 gives the similarity coefficient, as well as the number of domains in common, for every pair of methods. We observer that while there are some domains in common for any pair of methods, the ratio of the common domains is relatively small and the methods appear to be largely orthogonal to each other in the results they have produced. One exception to this is *Bing News Search API* and *Bing Search API Boolean*, which have 0.94 similarity coefficient. This likely indicates that *Bing News Search API* is primarily a repackaging of the main Bing Search Engine, but perhaps with slightly more focus on news sites.

6.2 Evaluation in Terms of Known Sites

Our first step assess the relevance and quality of domains discovered via each method, we cross check each set of results to see of how many of the known Fake News Domains are appear in the results of the discovery. We also check the number of known mainstream domains that appear in the discovery results. This number serves as a negative performance mark, in the sense that an ideal Fake News discovery method would yield all the Known Fake News Domains and none of the mainstream domains. Table 4 shows the numbers for each method. In this regard,we see that Alexa Overlap discovers the most of known fake news seeds (145). The three Bing API methods tried perform similarly well in number of KFNDs discovered. However, the results from these methods include even higher counts of known mainstream sites. We also note that the network methods yield no known mainstream sites, but they also discover the least number of KFNDs.

6.3 Classifier-aided Evaluation

While evaluating the results against the known ground truth can shed some light on the relative advantages of different methods, it does not address the problem of evaluating the thousands of unknow domains for which we have no ground truth. For this we conduct, one using a specialized classifier for fake news pages, and the other (described in the following section) using manual sampling.

The Fake News Classifier

In order to determine the quality and relevance of the discovery results, we use *TAG-FN* [5], a topic-agnostic classifier of fake news pages that uses linguistic and web markup features to predict the category as either "Fake News" or "Real News". *TAG-FN* was trained on a large corpus of content from known fake news and from known mainstream news domains. It achieves 0.87 F1 score in cross validation testing. One of the strengths of the classifier is that it does not use n-grams as features, but focuses on "topic-agnostic" style features. This should make to fake news sites that are focused on different themes and topics. For this evaluation of the discovery methods, we will treat the classifier as an oracle and assume its full accuracy.

Crawling Content from Discovered Domains

Since the classifier takes as input web articles not web sites as a whole, the first step of this evaluation required visiting the discovered domains and scraping content from them. We did this by effectively conducting a crawl of depth 1 on all the discovered sites, starting from their home pages. So for each discovered domain, we fetched the internal links on its homepage and then fetched the pages those links pointed to. One challenge here involved the substantial number of sites that could not be crawled (either because they had anti-scraping measures or because they had gone offline by the time we progressed to fetching content from them). Overall, we were able to get some content from 17,074 sites (82% of the total).

Filtering Crawled Content for English and Politics

The second step involved parsing the text out those pages and filtering for the pages that are (a) in English, and (b) involve Political topics. For the language detection we used the *lang-detect* library, while for the politics classification we trained a simple bag of n-grams classifier on the *HuffPost's News Category Dataset* [17]. This classifier achieved 0.88 accuracy on 5-fold cross-validation. The breakdown of filtering results can be seen in Table 5. Overall, around 62% of the discovered domains had majority English content, and 7% had Political content. However, since we are interested in finding fake news domains, even if they have non-Political content, we took any domain that had 10 or more Political pages as input to the TAG-FN classifier. There were 4080 domains (19% of the total) that met this criteria.

Classification Results

A comparison of the quality of each method's results, based on *TAG-FN's* predictions, can be seen in Table 6. We can see here that Bing news search API yielded the highest number of domains that could be crawled, and the highest number of domains that were classified as fake. However, ratio wise, the methods were comparable. Network-based methods did substantially better at filtering out non-English, non-Political sites but -in absolute numbers- this class of methods produced the lowest number of domains discovered and domains predicted fake news.

6.4 Evaluation by Manual Sampling

In addition to evaluating the discovery methods with the TAG FN Classifier, we conducted a manual evaluation using a random sample from each set of discovered domains. In comparison to the classifier-based evaluation, the manual evaluation offers some advantages and some disadvantages. In terms of the advantages of manual evaluation, the set of domains that we would be sampling

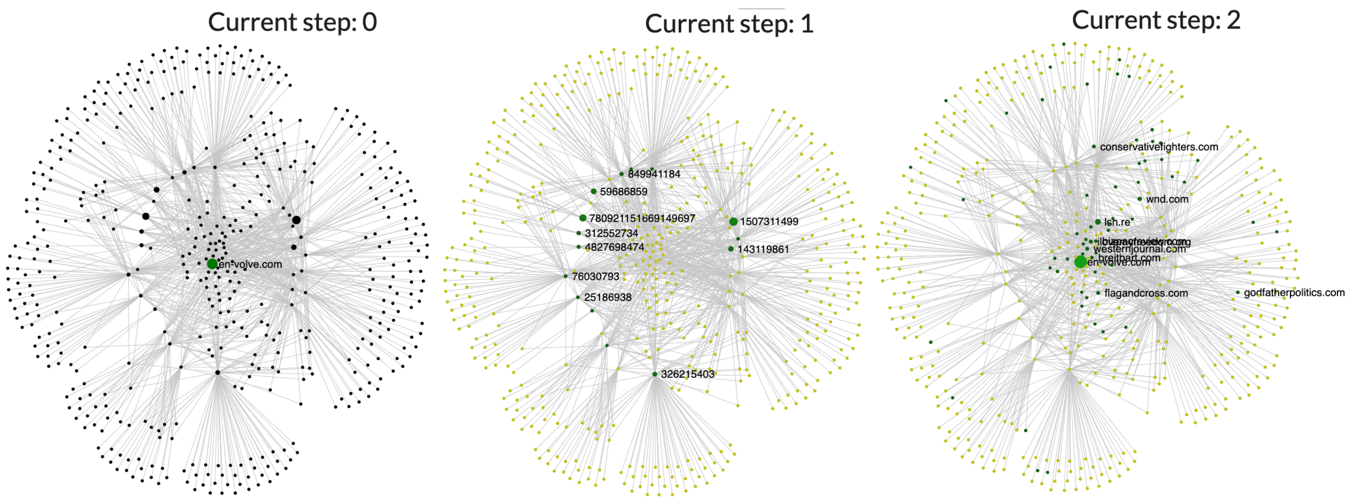

Figure 1: Illustration of two step random walk method. At step 0, we start with the seed node (in this case, the known fake news site is www.en-volve.com), giving the node a probability of 1. At step 1, the probability is proportionally distributed to users who tweet about this site. At step 2, probabilities flow back to domains from activated users (greens nodes in Step 1). We select top k domains sorted by probability in descending order. This method is effective at exploring graphs and discovering sites that are heavily tweeted by a single user. For example, conservativefighters.com and godfatherpolitics.com

	TNRW	TNJ	BNS	BSB	BSV	GSV	TCoS	AO	GR
GR	0.05 (93)	0.05 (75)	0.03 (458)	0.03 (456)	0.04 (314)	0.05 (161)	0.05 (192)	0.04 (195)	1.0 (1228)
AO	0.06 (244)	0.06 (216)	0.08 (1158)	0.08 (1156)	0.09 (850)	0.08 (437)	0.08 (456)	1.0 (3586)	
TCoS	0.1 (305)	0.09 (259)	0.07 (1000)	0.07 (993)	0.07 (631)	0.08 (344)	1.0 (2877)		
GSV	0.07 (179)	0.06 (133)	0.09 (1234)	0.09 (1237)	0.18 (1301)	1.0 (2004)			
BSV	0.04 (278)	0.03 (226)	0.28 (4273)	0.28 (4280)	1.0 (6641)				
BSB	0.03 (364)	0.02 (283)	0.94 (12496)	1.0 (12886)					
BNS	0.03 (367)	0.02 (284)	1.0 (12957)						
TNJ	0.3 (232)	1.0 (379)							
TNRW	1.0 (638)								

Table 3: Pairwise Jaccard Similarity Coefficient of Discovery Results, and the number of domains in common (in paratheses). Where GR is Google Related, AO is Alexa Overlap, TCoS is Twitter Co-Sharing, GSV is Google Search API - Verbatim, BSV is Bing Search API - Verbatim, BSB is Bing Search API - Boolean, BNS is Bing News Search API, TNJ is Twitter Network - Jaccard, TNRW is Twitter Network - Random Walk.

from is larger than the set of domains evaluated by the classifier since the problem of sites having anti-scraping or anti-crawling measures does not impede access. Further, while for the classifier based evaluation we limited ourselves to domains where 10 or more pages could be obtained, a manual labeling decision could be made based on the character of the site as a whole without needing to collect enough individual pages to meet a certain threshold. Additionally, the automated classification required the use of political classifier that, as it was trained on n-grams, often could not distinguish between political news sites and political advocacy sites. The manual labeling allows us to make that distinction and to focus only the news sites. The main disadvantage of manual sampling is perhaps obvious: we can only practically evaluate a small percentage of the discovered domains.

Sampling

We aimed to evaluate 5% from each set of discovered domains, while imposing a minimum of 20 and a maximum of 250 for any sampled set from one method. The sampling was random, and since the result sets are overlapping, the sampled sets from them are overlapping as well. In total, we manually labeled 554 domains.

Labeling Criteria

For each domain, we considered whether it meets the criteria for a Potential Fake News site: it needs to be a political site that presents itself as a news site or makes claims about current events. Additionally, it must have content that is either made-up or extremely

Discovery Method	Seed Type	Count of Domains Yielded	Count of KFNDs Discovered	Count of KMDs Discovered	Number of API Calls	Cost
Twitter Co-Sharing	Domain-based	2877	41	103	>10000	Free
Alexa Overlap	Domain-based	3586	145	45	~400	$149/month
Bing Search API- Verbatim	Content-based	6641	135	137	~3000	$12
Bing Search API- Boolean	Content-based	12886	140	206	~3000	$12
Bing News Search API	Content-based	12957	141	204	~3000	$12
Google Related	Domain-based	1228	40	45	~400	<$5
Google Search- Verbatim	Content-based	2004	108	52	~6000	$30
Twitter Network - Jaccard	Network-based	379	28	0	>10000	Free
Twitter Network - Random Walk	Network-based	638	36	0	>10000	Free

Table 4: Comparison of discovery methods by yield and cost

Discovery Method	Domains Discovered	Domains Not Crawled	Domains with 1-10 Pages Crawled	Domains with > 10 pages	Mostly English	Mostly Political	Domains with at least 10 English Poltical Pages
Google Related	1228	156 (12%)	128 (10%)	944 (76%)	887 (72%)	88 (7%)	284 (23%)
Alexa Overlap	3565	875 (24%)	367 (10%)	2323 (65%)	2084 (58%)	382 (10%)	819 (22%)
Twitter Co-Sharing	2691	526 (19%)	252 (9%)	1913 (71%)	1554 (57%)	316 (11%)	736 (27%)
Google Search API - Verbatim	2004	287 (14%)	112 (5%)	1605 (80%)	1370 (68%)	231 (11%)	613 (30%)
Bing Search API - Verbatim	6641	1396 (21%)	493 (7%)	4752 (71%)	4271 (64%)	600 (9%)	1543 (23%)
Bing Search API - Boolean	12886	1919 (14%)	811 (6%)	10156 (78%)	8840 (68%)	1098 (8%)	3010 (23%)
Bing News Search API	12957	1928 (14%)	812 (6%)	10217 (78%)	8881 (68%)	1105 (8%)	3023 (23%)
Twitter Network - Jaccard	379	26 (6%)	15 (3%)	338 (89%)	318 (83%)	119 (31%)	220 (58%)
Twitter Network-Random Walk	638	80 (12%)	38 (5%)	520 (81%)	467 (73%)	133 (20%)	289 (45%)
All Domain-Based Methods	6714	1486 (22%)	710 (10%)	4518 (67%)	3891 (57%)	580 (8%)	1378 (20%)
All Content-Based Methods	16227	2627 (16%)	1107 (6%)	12493 (76%)	10770 (66%)	1304 (8%)	3502 (21%)
All Network-Base Methods	785	92 (11%)	46 (5%)	647 (82%)	590 (75%)	177 (22%)	360 (45%)
ALL METHODS	20937	3863 (18%)	1714 (8%)	15360 (73%)	13168 (62%)	1578 (7%)	4080 (19%)

Table 5: Breakdown of the results of the crawling and filtering for the domains resulting from each method

distorted. Beside *Potential Fake News* label, we also label sites as *Likely Mainstream*, *Not working* for the sites that are unreachable or have a parking page, *Not-English* for websites that are not in English, and *Not Political News*. We report the manual evaluation results for each method and class of methods.

Observations from the Manual Evaluation Results

In considering the results of manual evaluation (Table 7), we note that the ratio of PFN domains is more varied between methods (and lower overall) than the ratios based on classifier predictions. We believe this is partly due to our filtering of political news sites from political advocacy sites (those belonging to a political organization or explicitly advancing a certain cause), a distinction that -as mentioned above- is difficult for an n-gram based politics classifier to make. We note that, similar to the Classifier-aided evaluation, network-based methods appear to be best at filtering out non-relevant sites (whether they are dormant, non-English , or non-politics domains). Qualitatively, the manual evaluation offered us an interesting window into the ecosystems and cultures that are related to fake news sites. Those ecosystems, which might be called "fake news-adjacent", involve many websites promoting spurious health claims or supplements, personal blogs of sometimes extreme nature, websites of various churches, sites of various political advocacy groups, etc. While it is beyond the scope of this work to analyze those sites (and since they did not meet our criteria, they were all labeled "Not Political News"), understanding how those domains exactly relate to the fake news domains (in terms of both audience and content) might be an essential step for understanding fake news phenomenon and is an interesting front for future research.

7 CONCLUSIONS AND FUTURE WORK

We have presented nine methods of varying complexity for discovering potential fake news domains on the web. We have applied those methods to a set of known fake news seeds. In the aggregate, the 9 discovery methods yielded over 20,000 domains. We

Discovered Method	Count of domains evaluated	Predicted fake (count)	Predicted fake (ratio)	Predicted real (count)	Predicted real (ratio)
Google Related	284	142	0.5	142	0.5
Alexa Overlap	819	509	0.62	310	0.38
Twitter Co-Sharing	736	401	0.54	335	0.46
Google Search- Verbatim	613	368	0.6	245	0.4
Bing Search API -Verbatim	1543	893	0.58	650	0.42
Bing Search API - Boolean	3010	1683	0.56	1327	0.44
Bing News Search API	3023	1680	0.56	1343	0.44
Twitter Network -Jaccard	220	125	0.57	95	0.43
Twitter Network -Random Walk	289	174	0.6	115	0.4
Domain-based Methods	1378	783	0.57	595	0.43
Content-based Methods	3502	1990	0.57	1512	0.43
Network-based Methods	360	214	0.59	146	0.41
ALL METHODS	4080	2339	0.57	1741	0.43

Table 6: Performance of each discovery method in terms of number of discovered domains and their classes, according to the TAG Fake News Classifier

Discovery Method	Discovered Domains	Manually Sampled	Not Working	Not English	Not Political News	Likely Mainstream	Potential Fake News
Google Related	1228	61	5 (8%)	4 (7%)	27 (44%)	11 (18%)	14 (23%)
Alexa Overlap	3565	178	33 (19%)	9 (5%)	58 (33%)	17 (10%)	61 (34%)
Twitter Co-Sharing	2691	134	12 (9%)	17 (13%)	53 (40%)	17 (13%)	35 (26%)
Google Search API - Verbatim	2004	100	6 (6%)	5 (5%)	42 (42%)	6 (6%)	41 (41%)
Bing Search API - Verbatim	6641	250	34 (14%)	16 (6%)	101 (40%)	28 (11%)	71 (28%)
Bing Search API - Boolean	12886	250	27 (11%)	17 (7%)	103 (41%)	27 (11%)	76 (30%)
Bing News Search API	12957	250	22 (9%)	14 (6%)	111 (44%)	32 (13%)	71 (28%)
Twitter Network - Jaccard	379	20	1 (5%)	0 (0%)	10 (50%)	0 (0%)	9 (45%)
Twitter Network - Random Walk	638	31	2 (6%)	0 (0%)	8 (26%)	3 (10%)	18 (58%)
All Domain-based Methods	6714	299	45 (15%)	30 (10%)	120 (40%)	33 (11%)	71 (24%)
All Content-based Methods	16227	394	41 (10%)	31 (8%)	175 (44%)	44 (11%)	103 (26%)
All Network-based Methods	785	47	2 (4%)	0 (0%)	16 (34%)	3 (6%)	26 (55%)
ALL METHODS	20937	554	74 (13%)	58 (10%)	247 (45%)	56 (10%)	119 (21%)

Table 7: Results of manual evaluation of a random sample from each method's discovered domains. The sampled sets are overlapping

conducted a classifier-based evaluation of the results as well as a manual labeling of random samples. Our results indicate that there are substantial numbers of discovered domains unique to each methods. Our classifier-based evaluation indicated comparable rates of Potential Fake News Domains among the methods, with network-based yielding a higher ratio of relevant domains but fewer in absolute numbers. Our manual evaluation indicated lower, but still substantial, rates of potential fake news domains in the discovered results. Overall, our evaluation shows that each of these methods is individually useful to the task of identifying candidate fake news sites, and -given the relatively small overlap- an ensemble of all such methods is needed for a comprehensive approach.

For future work, we believe there are worthwhile opportunities offered by the large data set of potential Fake News sites yielded by the discovery methods. Beyond being a necessary component to a web-centric fake news Classification system, it can allow a fine grained analysis of how certain fake news spread on the web on social media, i.e. the dynamics, speed of the propagation, and which nodes are central to the process. Additionally, there is an important question about the interplay between these fake news sites and the partisan mainstream news sites. We saw glimpses of this relation in our evaluation in the fact that some mainstream sites kept being discovered from known fake news seeds. Studying this flow of audience and content between these two classes of sites would go a long way towards a better understanding of how the spread of fake news is enabled by established media. Finally, we focused our attention in this work on the class of political fake news sites. However, we saw in the course of our qualitative evaluation numerous sites that are promoting questionable non-political content. Such content included unproven health claims, various fringe economic and religious viewpoints, claims about supernatural phenomena, etc. A deeper study of these types of content might help further the understanding of the fake news ecosystem and its place within the web at large.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31, 2 (2017), 211–36.

[2] Luciano Barbosa. 2017. Harvesting Forum Pages from Seed Sites. In *Web Engineering*, Jordi Cabot, Roberto De Virgilio, and Riccardo Torlone (Eds.). Springer International Publishing, 457–468.

[3] Luciano Barbosa, Srinivas Bangalore, and Vivek Kumar Rangarajan Sridhar. 2011. Crawling Back and Forth: Using Back and Out Links to Locate Bilingual Sites. In *Proceedings of the 5th International Joint Conference on Natural Language Processing, IJCNLP*. 429–437.

[4] BuzzFeedNews. [n. d.]. Analysis of fake news sites and viral posts, 2016 vs. 2017. https://github.com/BuzzFeedNews/2017-12-fake-news-top-50.

[5] Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. 2019. A Topic-Agnostic Approach for Identifying Fake News Pages. In *Companion Proceedings of The 2019 World Wide Web Conference (WWW '19)*. ACM, New York, NY, USA, 975–980.

[6] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. 675–684.

[7] Zhouhan Chen and Devika Subramanian. 2018. An Unsupervised Approach to Detect Spam Campaigns that Use Botnets on Twitter. arXiv:arXiv:1804.05232

[8] Camila Domonoske. 2016. Students Have Dismaying Inability To Tell Fake News From Real, Study Finds.

[9] Brett Edkins. 2016. Americans Believe They Can Detect Fake News. Studies Show they Can't. www.forbes.com/sites/brettedkins/2016/12/20/americans-believe-they-can-detect-fake-news-studies-show-they-cant/.

[10] James Fairbanks, Natalie Fitch, Nathan Knauf, and Erica Briscoe. 2018. Credibility Assessment in the News: Do we need to read?. In *Proceedings of the MIS2 Workshop held in conjuction with 11th International Conference on Web Search and Data Mining*. 799–800.

[11] Joshua Gillin. 2017. PolitiFact's guide to fake news websites and what they peddle. https://www.politifact.com/punditfact/article/2017/apr/20/politifacts-guide-fake-news-websites-and-what-they/.

[12] Benjamin D Horne and Sibel Adali. 2017. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the 2nd International Workshop on News and Public Opinion*.

[13] Seyedmehdi Hosseinimotlagh and Evangelos E. Papalexakis. 2018. Unsupervised Content-Based Identification of Fake News Articles with Tensor Decomposition Ensembles. In *Proceedings of the MIS2 Workshop held in conjuction with 11th International Conference on Web Search and Data Mining*. 799–800.

[14] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[15] David Liben-Nowell and Jon Kleinberg. 2003. The Link Prediction Problem for Social Networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM '03)*. ACM, New York, NY, USA, 556–559.

[16] Annie Louis and Ani Nenkova. 2013. What makes writing great? First experiments on article quality prediction in the science journalism domain. *Transactions of the Association for Computational Linguistics* 1 (2013), 341–352.

[17] Rishabh Misra. 2018. News Category Dataset. https://doi.org/10.13140/RG.2.2.20331.18729

[18] Tsuyoshi Murata. 2001. Finding Related Web Pages Based on Connectivity Information from a Search Engine. In *Poster Proceedings of 10th International Conference on World Wide Web (WWW)*. 18–19.

[19] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *International Conference on Computational Linguistics*. 3391–3401.

[20] Kien Pham, Aécio Santos, and Juliana Freire. 2018. Learning to Discover Domain-Specific Web Content. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 432–440.

[21] Kien Pham, Aecio Santos, and Juliana Freire. 2019. Bootstrapping Domain-Specific Content Discovery on the Web. In *The World Wide Web Conference (WWW '19)*. ACM, New York, NY, USA, 1476–1486.

[22] Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 186–195.

[23] A. A. Putra, R. Mahendra, I. Budi, and Q. Munajat. 2017. Two-steps graph-based collaborative filtering using user and item similarities: Case study of E-commerce recommender systems. In *2017 International Conference on Data and Software Engineering (ICoDSE)*. 1–6.

[24] Disheng Qiu, Luciano Barbosa, Xin Luna Dong, Yanyan Shen, and Divesh Srivastava. 2015. Dexter: Large-scale Discovery and Extraction of Product Specifications on the Web. *Proc. VLDB Endow.* 8, 13 (Sept. 2015), 2194–2205.

[25] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2931–2937.

[26] Sujata A. Sawant. 2013. Collaborative Filtering using Weighted BiPartite Graph Projection - A Recommendation System for Yelp. In *Proceedings of the CS224W: Social and Information Network Analysis Conference*.

[27] Kai Shu, Suhang Wang, and Huan Liu. 2017. Exploiting Tri-Relationship for Fake News Detection. *CoRR* abs1712.07709 (2017).

[28] Craig Silverman, Jane Lytvynenko, and Scott Pham. 2017. These Are 50 Of The Biggest Fake News Hits On Facebook In 2017. https://www.buzzfeednews.com/article/craigsilverman/these-are-50-of-the-biggest-fake-news-hits-on-facebook-in

[29] Karane Vieira, Luciano Barbosa, Altigran Soares da Silva, Juliana Freire, and Edleno Moura. 2016. Finding seeds to bootstrap focused crawlers. *World Wide Web* 19, 3 (2016), 449–474.

[30] Xuezhi Wang, Cong Yu, Simon Baumgartner, and Flip Korn. 2018. Relevant Document Discovery for Fact-Checking Articles. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 525–533.

A APPENDIX - DISCOVERY SEEDS

1) breitbart.com [OpenSources, Hoaxy]
2) infowars.com [OpenSources, Hoaxy, Wikipedia]
3) wnd.com [OpenSources, Hoaxy]
4) naturalnews.com [OpenSources, Hoaxy, Wikipedia]
5) yournewswire.com [Poilitfact, Hoaxy, BuzzFeed, OpenSources, Wikipedia]
6) worldtruth.tv [Poilitfact, OpenSources, Hoaxy, Wikipedia]
7) prisonplanet.com [OpenSources, Hoaxy]
8) redstate.com [OpenSources, Hoaxy]
9) activistpost.com [OpenSources, Hoaxy]
10) thedailysheeple.com [OpenSources, Hoaxy]
11) thefreethoughtproject.com [OpenSources, Hoaxy]
12) conservativedailypost.com [Poilitfact, OpenSources]
13) twitchy.com [OpenSources, Hoaxy]
14) onepoliticalplaza.com [Poilitfact, OpenSources]
15) veteranstoday.com [OpenSources, Hoaxy]
16) theblaze.com [OpenSources, Hoaxy]
17) clashdaily.com [Poilitfact, OpenSources]
18) madworldnews.com [Poilitfact, OpenSources]
19) huzlers.com [Poilitfact, OpenSources, BuzzFeed, Wikipedia]
20) adobochronicles.com [OpenSources, BuzzFeed]
21) 21stcenturywire.com [OpenSources, Hoaxy]
22) bipartisanreport.com [OpenSources, Hoaxy]
23) disclose.tv [Poilitfact, OpenSources, Hoaxy, Wikipedia]
24) conservative101.com [Poilitfact, Wikipedia]
25) hangthebankers.com [OpenSources, Hoaxy]
26) bb4sp.com [Poilitfact, OpenSources]
27) worldnewsdailyreport.com [Poilitfact, Hoaxy, BuzzFeed, OpenSources, Wikipedia]
28) whatdoesitmean.com [Poilitfact, OpenSources]
29) govtslaves.info [OpenSources, Hoaxy]
30) bizstandardnews.com [BuzzFeed, Wikipedia]
31) thevalleyreport.com [OpenSources, BuzzFeed]
32) intellihub.com [OpenSources, Hoaxy]
33) occupydemocrats.com [OpenSources, Hoaxy]
34) neonnettle.com [Poilitfact, OpenSources, BuzzFeed]
35) lewrockwell.com [OpenSources, Hoaxy]
36) weeklyworldnews.com [OpenSources, Hoaxy]
37) freedomsfinalstand.com [Poilitfact]
38) nowtheendbegins.com [OpenSources, Hoaxy]
39) realfarmacy.com [OpenSources, Hoaxy]
40) realnewsrightnow.com [Poilitfact, OpenSources, BuzzFeed]
41) democraticunderground.com [Poilitfact]
42) notallowedto.com [Poilitfact, OpenSources, BuzzFeed]
43) rearfront.com [Poilitfact]
44) americannewsx.com [OpenSources, Hoaxy]
45) now8news.com [Poilitfact, OpenSources, BuzzFeed, Wikipedia]
46) thespoof.com [OpenSources, Hoaxy]
47) christiantimes.com [Hoaxy]
48) patriotcrier.com [Poilitfact]
49) consciouslifenews.com [OpenSources, Hoaxy]
50) gulagbound.com [OpenSources, Hoaxy]
51) angrypatriotmovement.com [Poilitfact, OpenSources]
52) ifyouonlynews.com [OpenSources, Hoaxy]
53) politicususa.com [OpenSources, Hoaxy]
54) coasttocoastam.com [OpenSources, Hoaxy]
55) beforeitsnews.com [Poilitfact, OpenSources, Hoaxy, Wikipedia]
56) fprnradio.com [OpenSources, Hoaxy]
57) newsbiscuit.com [OpenSources, BuzzFeed]
58) burrardstreetjournal.com [OpenSources, BuzzFeed, Hoaxy]
59) americannews.com [Poilitfact, OpenSources, Hoaxy, Wikipedia]
60) newsbreakshere.com [Poilitfact, OpenSources, Wikipedia]
61) ustruthwire.com [Poilitfact]
62) givemeliberty01.com [Poilitfact]
63) usherald.com [Poilitfact, OpenSources]
64) tmzbreaking.com [BuzzFeed]
65) dailybuzzlive.com [Poilitfact, OpenSources, Wikipedia]
66) celebtricity.com [OpenSources, BuzzFeed, Wikipedia]
67) viralactions.com [Poilitfact, OpenSources, BuzzFeed]
68) politicalblindspot.com [OpenSources, Hoaxy]
69) ladylibertysnews.com [Poilitfact]
70) news4ktla.com [Poilitfact, OpenSources, BuzzFeed]
71) freewoodpost.com [Poilitfact, OpenSources, BuzzFeed, Hoaxy]
72) en-volve.com [Poilitfact]
73) themoralofthestory.us [Poilitfact, OpenSources]
74) 12minutos.com [BuzzFeed]
75) usasupreme.com [Poilitfact, OpenSources]
76) viralspeech.com [BuzzFeed]
77) stuppid.com [OpenSources, BuzzFeed, Wikipedia]
78) globalresearch.ca [OpenSources, Hoaxy]
79) dcclothesline.com [OpenSources, Hoaxy]
80) liberalamerica.org [OpenSources, Hoaxy]
81) actualite.co [BuzzFeed]
82) badcriminals.com [BuzzFeed]
83) 24aktuelles.com [BuzzFeed]
84) noticias-frescas.com [BuzzFeed]
85) smhwtfnews.com [BuzzFeed]
86) therundownlive.com [Hoaxy]
87) theinternetpost.net [Poilitfact, OpenSources]
88) firebrandleft.com [Wikipedia]
89) newsuptoday.com [Poilitfact]
90) libertytalk.fm [OpenSources, Hoaxy]
91) dcgazette.com [OpenSources, Wikipedia]
92) breakingnewsblast.com [Poilitfact]
93) channel45news.com [BuzzFeed]
94) viraldevil.com [BuzzFeed]
95) journaldemourreal.com [BuzzFeed]
96) folksvideo.com [BuzzFeed]
97) empirenews.net [Poilitfact, OpenSources, BuzzFeed, Wikipedia]
98) viralmugshot.com [BuzzFeed]
99) cartelreport.com [BuzzFeed]
100) react365.com [Poilitfact, OpenSources, BuzzFeed, Wikipedia]
101) channel28news.com [BuzzFeed]
102) theneonnettle.com [Wikipedia]
103) dailysurge.com [Poilitfact, OpenSources]
104) 20minutenews.com [BuzzFeed]
105) thefrt.com [BuzzFeed]
106) politicalmayhem.news [Poilitfact]
107) thenet24h.com [Poilitfact, OpenSources]
108) newslo.com [Poilitfact, OpenSources, BuzzFeed]
109) freedomdaily.com [Poilitfact, OpenSources]
110) channel23news.com [BuzzFeed]

111) tmzcomedy.com [BuzzFeed]
112) democraticmoms.com [Poilitfact]
113) puppetstringnews.com [Poilitfact]
114) qualitysharing.com [BuzzFeed]
115) redrocktribune.com [Poilitfact, OpenSources]
116) thelastlineofdefense.org [Poilitfact, OpenSources, BuzzFeed, Wikipedia]
117) dailycurrant.com [OpenSources, BuzzFeed]
118) president45donaldtrump.com [Poilitfact, OpenSources]
119) areyousleep.com [BuzzFeed]
120) reaganwasright.com [BuzzFeed]
121) federalisttribune.com [Poilitfact]
122) usapoliticstoday.com [Poilitfact, OpenSources]
123) nationalreport.net [Poilitfact, Hoaxy, BuzzFeed, OpenSources, Wikipedia]
124) mediamass.net [OpenSources, BuzzFeed]
125) thenochill.com [BuzzFeed]
126) viralpropaganda.com [BuzzFeed]
127) socialeverythings.com [Poilitfact, OpenSources]
128) en-bref.fr [BuzzFeed]
129) nnettle.com [Poilitfact]
130) breakingnews365.net [Poilitfact, BuzzFeed]
131) thenewyorkevening.com [Poilitfact, OpenSources, BuzzFeed, Wikipedia]
132) washingtonfeed.com [Poilitfact]
133) 70news.wordpress.com [OpenSources, Hoaxy, Wikipedia]
134) drugsofficial.com [BuzzFeed]
135) 16wmpo.com [Poilitfact, OpenSources]
136) pakalertpress.com [OpenSources, Hoaxy]
137) ncscooper.com [OpenSources, BuzzFeed, Hoaxy]
138) flashnewscorner.com [Poilitfact, OpenSources]
139) theracketreport.com [Poilitfact, OpenSources, BuzzFeed]
140) channel24news.com [Poilitfact, BuzzFeed]
141) endingthefed.com [OpenSources, Hoaxy]
142) anonews.co [OpenSources, Hoaxy]
143) nachrichten365.com [BuzzFeed]
144) defenseusa.club [Poilitfact]
145) guerilla.news [Poilitfact]
146) newzmagazine.com [Poilitfact]
147) usadosenews.com [Poilitfact, OpenSources]
148) uspoliticsinfo.com [Poilitfact]
149) westfieldpost.com [Poilitfact]
150) empiresports.co [Poilitfact, OpenSources, BuzzFeed, Wikipedia]
151) politicops.com [Poilitfact, OpenSources, BuzzFeed]
152) redflagnews.com [OpenSources, Hoaxy, Wikipedia]
153) channel22news.com [BuzzFeed]
154) thelastlineofdefense.online [BuzzFeed]
155) americanflavor.news [Poilitfact, BuzzFeed]
156) departedme.com [Poilitfact]
157) nativestuff.us [Poilitfact]
158) newsexaminer.net [Poilitfact, OpenSources, BuzzFeed]
159) newsfeedhunter.com [Poilitfact, BuzzFeed]
160) rogue-nation3.com [Poilitfact, OpenSources]
161) usanewsflash.com [Poilitfact, OpenSources]
162) wazanews.tk [Poilitfact]
163) observeronline.news [Poilitfact]
164) wrpm33.com [Poilitfact]
165) winningdemocrats.com [OpenSources, Hoaxy]
166) witscience.org [OpenSources, BuzzFeed]
167) gossipmillsa.com [Wikipedia]
168) civictribune.com [Poilitfact, OpenSources, BuzzFeed]
169) newsoftrump.com [Poilitfact]
170) tdtalliance.com [Poilitfact, OpenSources]
171) theusaconservative.com [Poilitfact]
172) usadailytime.com [Poilitfact, OpenSources, BuzzFeed]
173) usinfonews.com [Poilitfact, OpenSources]
174) uspoln.com [Poilitfact, OpenSources, BuzzFeed]
175) breakingnews247.net [Poilitfact, BuzzFeed]
176) dailynews10.com [Poilitfact, OpenSources]
177) kcst7.com [Poilitfact]
178) newsdaily12.com [Poilitfact, OpenSources]
179) thewashingtonpress.com [Poilitfact, OpenSources]
180) politicot.com [Poilitfact, OpenSources, BuzzFeed]
181) rilenews.com [OpenSources, BuzzFeed]
182) thenewsnerd.com [OpenSources, BuzzFeed, Wikipedia]
183) dailyfinesser.com [BuzzFeed]
184) newspunch.com [BuzzFeed]
185) pressunion.org [BuzzFeed]
186) thebreakingnews.co [BuzzFeed]
187) 247newsmedia.com [Poilitfact]
188) americantoday.us [Poilitfact]
189) 24x365live.com [Poilitfact, BuzzFeed]
190) globalpoliticsnow.com [Poilitfact, OpenSources]
191) liberalplug.com [Poilitfact, OpenSources]
192) newsconservative.com [Poilitfact, BuzzFeed]
193) newsjustforyou1.blogspot.com [Poilitfact]
194) openmagazines.com [Poilitfact, OpenSources]
195) proudleader.com [Poilitfact]
196) religionmind.com [Poilitfact, OpenSources]
197) revolutions2040.com [Poilitfact, OpenSources]
198) thirdestatenewsgroup.com [Poilitfact]
199) usanews4u.us [Poilitfact]
200) wetheproudpatriots.com [Poilitfact]
201) kbc14.com [Poilitfact]
202) local31news.com [Poilitfact, OpenSources]
203) wy21news.com [Poilitfact]
204) alertchild.com [BuzzFeed]
205) channel34news.com [BuzzFeed]
206) channel46news.com [BuzzFeed]
207) unitednews.org [BuzzFeed]
208) liberalsociety.com [Wikipedia]
209) armyusanews.com [Poilitfact]
210) dailyfeed.news [Poilitfact]
211) donaldtrumppotus45.com [Poilitfact, OpenSources]
212) libertyalliance.com [Poilitfact, OpenSources]
213) metropolitanworlds.com [Poilitfact, OpenSources]
214) nephef.com [Poilitfact, OpenSources]
215) prntly.com [Poilitfact, OpenSources, Wikipedia]
216) redcountry.us [Poilitfact, OpenSources, BuzzFeed]
217) redpolitics.us [Poilitfact]
218) success-street.com [Poilitfact, OpenSources]
219) thetrumpmedia.com [Poilitfact, OpenSources]
220) usapoliticsnow.com [Poilitfact, OpenSources]
221) usapoliticszone.com [Poilitfact, OpenSources]

222) ushealthyadvisor.com [Poilitfact, OpenSources]
223) worldpoliticsnow.com [Poilitfact, OpenSources]
224) abcnews.com.co [Poilitfact, OpenSources, BuzzFeed, Wikipedia]
225) christiantimesnewspaper.com [Poilitfact, OpenSources, BuzzFeed]
226) denverinquirer.com [Poilitfact, BuzzFeed]
227) empireherald.com [Poilitfact, OpenSources, BuzzFeed, Hoaxy]
228) fox-news24.com [Poilitfact]
229) houstonchronicle-tv.com [Poilitfact, BuzzFeed]
230) usa-television.com [Poilitfact, OpenSources]
231) floridasunpost.com [Poilitfact, OpenSources, BuzzFeed]
232) theseattletribune.com [Poilitfact, OpenSources, BuzzFeed]
233) knp7.com [BuzzFeed]
234) straightstoned.com [BuzzFeed]
235) abcnews-us.com [BuzzFeed]
236) cnoticias.net [BuzzFeed]
237) dailyworldupdate.com [BuzzFeed]
238) mrnewswatch.com [BuzzFeed]
239) nachrichten.de.com [BuzzFeed]
240) toutelinfo.fr [BuzzFeed]
241) lightlybraisedturnip.com [Hoaxy]
242) linkbeef.com [Wikipedia]
243) a-news24.com [Poilitfact]
244) americanpeoplenetwork.com [Poilitfact]
245) americanpoliticnews.co [Poilitfact, OpenSources]
246) americanpresident.co [Poilitfact, OpenSources]
247) amposts.com [Poilitfact, OpenSources]
248) anonjekloy.tk [Poilitfact, OpenSources]
249) antinews.com [Poilitfact]
250) bluevision.news [Poilitfact]
251) bluevisionpost.com [Poilitfact]
252) breakingtop.world [Poilitfact]
253) buzzfeedusa.com [Poilitfact, OpenSources]
254) christiantoday.info [Poilitfact]
255) cnewsgo.com [Poilitfact]
256) coffeebreakforyou.com [Poilitfact]
257) conservativearmy88.com [Poilitfact]
258) 24wpn.com [Poilitfact, OpenSources]
259) conservativeflashnews.com [Poilitfact, OpenSources]
260) conservativeinfocorner.com [Poilitfact]
261) conservativepaper.com [Poilitfact]
262) conservativespirit.com [Poilitfact, OpenSources]
263) conservativeview.info [Poilitfact]
264) dailyinfobox.com [Poilitfact, OpenSources]
265) dailynewsposts.info [Poilitfact, OpenSources]
266) dailysidnews.com [Poilitfact]
267) dailythings.world [Poilitfact]
268) dailyusaupdate.com [Poilitfact, OpenSources]
269) damnleaks.com [Poilitfact, OpenSources]
270) departedmedia.com [Poilitfact, OpenSources]
271) diyhours.net [Poilitfact, OpenSources]
272) enhlive.com [Poilitfact, OpenSources]
273) fedsalert.com [Poilitfact, OpenSources, BuzzFeed]
274) focusnews.info [Poilitfact]
275) forfreedomworld.com [Poilitfact]
276) freeinfomedia.com [Poilitfact]
277) freshdailyreport.com [Poilitfact, OpenSources]
278) globalrevolutionnetwork.com [Poilitfact]
279) gummypost.com [Poilitfact, OpenSources, Wikipedia]
280) healthyworldhouse.com [Poilitfact, OpenSources]
281) interestingdailynews.com [Poilitfact, OpenSources]
282) isthatlegit.com [Poilitfact]
283) konkonsagh.biz [Poilitfact, OpenSources]
284) lastdeplorables.com [Poilitfact, OpenSources]
285) liberty-courier.com [Poilitfact]
286) majorthoughts.com [Poilitfact, OpenSources]
287) newpoliticstoday.com [Poilitfact, OpenSources]
288) newsbreakhere.com [Poilitfact]
289) newsbreakingspipe.com [Poilitfact]
290) newsbysquad.com [Poilitfact, OpenSources]
291) newsformetoday.com [Poilitfact]
292) newsleak.co [Poilitfact, OpenSources]
293) newsphd.com [Poilitfact]
294) patriotusa.website [Poilitfact]
295) persecutes.com [Poilitfact, BuzzFeed]
296) politicspaper.com [Poilitfact]
297) politicsusanews.com [Poilitfact, OpenSources]
298) readconservatives.news [Poilitfact, OpenSources]
299) redinfo.us [Poilitfact, BuzzFeed]
300) rumorjournal.com [Poilitfact, OpenSources]
301) snoopack.com [Poilitfact]
302) southernconservativeextra.com [Poilitfact]
303) spinzon.com [Poilitfact]
304) states-tv.com [Poilitfact]
305) stgeorgegazette.com [Poilitfact]
306) supremepatriot.com [Poilitfact, OpenSources]
307) teoinfo.com [Poilitfact]
308) the-insider.co [Poilitfact, OpenSources]
309) thepremiumnews.com [Poilitfact]
310) theexaminer.site [Poilitfact]
311) the-global-news.com [Poilitfact]
312) theusa-news.com [Poilitfact, OpenSources]
313) theworldupdate.com [Poilitfact]
314) tmzworldnews.com [Poilitfact, OpenSources, BuzzFeed]
315) trueamericans.me [Poilitfact, OpenSources]
316) truetrumpers.com [Poilitfact, OpenSources, Wikipedia]
317) undergroundnewsreport.com [Poilitfact, OpenSources, BuzzFeed, Wikipedia]
318) universepolitics.com [Poilitfact, OpenSources]
319) usa360-tv.com [Poilitfact]
320) usa-radio.com [Poilitfact, OpenSources]
321) usadailyinfo.com [Poilitfact, OpenSources]
322) usadailypost.us [Poilitfact]
323) usafirstinformation.com [Poilitfact, OpenSources]
324) usanewstoday.com [Poilitfact, OpenSources]
325) usapolitics24hrs.com [Poilitfact, OpenSources]
326) usasnich.com [Poilitfact, OpenSources]
327) usaworldbox.com [Poilitfact]
328) ushealthylife.com [Poilitfact, OpenSources]
329) usadailythings24.com [Poilitfact]
330) usanewshome.com [Poilitfact]
331) uspostman.com [Poilitfact, OpenSources]
332) usaconservativereport.com [Poilitfact]

333) veteransfordonaldtrump.com [Poilitfact]
334) voxtribune.com [Poilitfact, OpenSources]
335) weconservative.com [Poilitfact, BuzzFeed]
336) weekendpoliticalnews.com [Poilitfact]
337) worldnewscircle.com [Poilitfact]
338) world.politics.com [Poilitfact]
339) 24online.news [Poilitfact]
340) alabamaobserver.com [Poilitfact, BuzzFeed]
341) aurora-news.us [Poilitfact, OpenSources]
342) channel18news.com [Poilitfact, OpenSources, BuzzFeed]
343) cnn-business-news.ga [Poilitfact]
344) cnnews3.com [Poilitfact, OpenSources]
345) dailynews5.com [Poilitfact, OpenSources, BuzzFeed]
346) kf13.com [Poilitfact]
347) kmt11.com [Poilitfact, OpenSources, BuzzFeed, Wikipedia]
348) krb7.com [Poilitfact]
349) kty24news.com [Poilitfact]
350) ky12news.com [Poilitfact, OpenSources, BuzzFeed]
351) mississippiherald.com [Poilitfact, BuzzFeed]
352) nbc.com.co [Poilitfact, OpenSources]
353) nydailynews-tv.com [Poilitfact]
354) themiamigazette.com [Poilitfact]
355) times.com.mx [Poilitfact, OpenSources]
356) usatodaynews.me [Poilitfact, OpenSources]
357) washingtonevening.com [Poilitfact, BuzzFeed]
358) washingtonpost.com.co [Poilitfact, OpenSources, Wikipedia]
359) werk35.com [Poilitfact]
360) associatedmediacoverage.com [Poilitfact, OpenSources, BuzzFeed]
361) bostonleader.com [Poilitfact, OpenSources, BuzzFeed]
362) satiratribune.com [Poilitfact, OpenSources, BuzzFeed]
363) nymeta.co [OpenSources, BuzzFeed]
364) usuncut.com [OpenSources, Hoaxy]
365) denverguardian.com [OpenSources, BuzzFeed, Wikipedia]
366) drudgereport.com.co [OpenSources, Wikipedia]
367) krbcnews.com [OpenSources, BuzzFeed]
368) msnbc.website [OpenSources, BuzzFeed]
369) newswatch33.com [OpenSources, BuzzFeed]
370) thebostontribune.com [OpenSources, BuzzFeed, Wikipedia]
371) thereporterz.com [OpenSources, BuzzFeed]
372) unitedmediapublishing.com [OpenSources, BuzzFeed, Wikipedia]
373) usatoday.com.co [OpenSources, Wikipedia]
374) wtoe5news.com [OpenSources, BuzzFeed]
375) alynews.com [BuzzFeed]
376) baldwinpost.com [BuzzFeed]
377) baltimoregazette.com [BuzzFeed]
378) cartelpress.com [BuzzFeed]
379) channel16news.com [BuzzFeed]
380) channel17news.com [BuzzFeed]
381) clancyreport.com [BuzzFeed]
382) dailynews11.com [BuzzFeed]
383) dailynews3.com [BuzzFeed]
384) dailynews33.com [BuzzFeed]
385) freddymag.com [BuzzFeed]
386) globalassociatednews.com [BuzzFeed]
387) headlinebrief.com [BuzzFeed]
388) heaviermetal.net [BuzzFeed]
389) hotglobalnews.com [BuzzFeed]
390) kata33.com [BuzzFeed]
391) klponews.com [BuzzFeed]
392) km8news.com [BuzzFeed]
393) kspm33.com [BuzzFeed]
394) kupr7.com [BuzzFeed]
395) ky6news.com [BuzzFeed]
396) kypo6.com [BuzzFeed]
397) landrypost.com [BuzzFeed]
398) localnews33.com [BuzzFeed]
399) lopezreport.com [BuzzFeed]
400) maywoodpost.com [BuzzFeed]
401) mbynews.com [BuzzFeed]
402) mckenziepost.com [BuzzFeed]
403) news14kgpn.com [BuzzFeed]
404) news14now.com [BuzzFeed]
405) newsbuzzdaily.com [BuzzFeed]
406) newsdaily10.com [BuzzFeed]
407) newsdaily27.com [BuzzFeed]
408) newshubs.info [BuzzFeed]
409) newsnow17.com [BuzzFeed]
410) oreillypost.com [BuzzFeed]
411) scrapetv.com [BuzzFeed]
412) thenationalsun.com [BuzzFeed]
413) tmzhiphop.com [BuzzFeed]
414) wcpm3.com [BuzzFeed]
415) wftj8news.com [BuzzFeed]
416) wleb21.com [BuzzFeed]
417) wm21news.com [BuzzFeed]
418) wmacnews.com [BuzzFeed]
419) wmb36.com [BuzzFeed]
420) wrejnews.com [BuzzFeed]
421) wrpt16.com [BuzzFeed]
422) albertatimes.com [BuzzFeed]
423) battypost.com [BuzzFeed]
424) borderherald.com [BuzzFeed]
425) channel33news.com [BuzzFeed]
426) channel40news.com [BuzzFeed]
427) channel5000.com [BuzzFeed]
428) channel55news.com [BuzzFeed]
429) channel56news.com [BuzzFeed]
430) channel59news.com [BuzzFeed]
431) channel60news.com [BuzzFeed]
432) channel62news.com [BuzzFeed]
433) channel63news.com [BuzzFeed]
434) channel65news.com [BuzzFeed]
435) channel66news.com [BuzzFeed]
436) channel68news.com [BuzzFeed]
437) channel77news.com [BuzzFeed]
438) cnn-globalnews.com [BuzzFeed]
439) cnn-internationaledition.com [BuzzFeed]
440) cnn.com.de [BuzzFeed]
441) countyweekly.com [BuzzFeed]
442) dallastimes.us [BuzzFeed]
443) dublintribune.com [BuzzFeed]
444) everydaybreakingnews.com [BuzzFeed]
445) freeinquirer.com [BuzzFeed]

446) `houstonleader.com` [BuzzFeed]
447) `jacksontelegraph.com` [BuzzFeed]
448) `jokerviral.com` [BuzzFeed]
449) `ladiesofliberty.net` [BuzzFeed]
450) `net-infosnews.com` [BuzzFeed]
451) `newjournal.us` [BuzzFeed]
452) `notizzia.com` [BuzzFeed]
453) `potatriotpost.com` [BuzzFeed]
454) `scaryhours.com` [BuzzFeed]
455) `sundayinquirer.com` [BuzzFeed]
456) `sundaypost.org` [BuzzFeed]
457) `superstation95.com` [BuzzFeed]
458) `surreytelegraph.com` [BuzzFeed]
459) `sydneybulletin.com` [BuzzFeed]
460) `telegraphsun.com` [BuzzFeed]
461) `timesofcambodia.com` [BuzzFeed]
462) `tmzuncut.com` [BuzzFeed]
463) `tmzworldstar.com` [BuzzFeed]
464) `tmzworldstarnews.com` [BuzzFeed]
465) `vancouverinquirer.com` [BuzzFeed]
466) `viralstupid.com` [BuzzFeed]
467) `weekendherald.com` [BuzzFeed]
468) `en.mediamass.net` [Hoaxy]
469) `thedcgazette.com` [Hoaxy]
470) `worldnewsandpolitics.wordpress.com` [Wikipedia]
471) `countynewsroom.info` [Wikipedia]
472) `bloomberg.ma` [Wikipedia]
473) `cnn-trending.com` [Wikipedia]
474) `conservativefrontline.com` [Wikipedia]
475) `currentish.com` [Wikipedia]
476) `ww7.libertywritersnews.com` [Wikipedia]
477) `nbcnews.com.co` [Wikipedia]