Lay it Out: Detecting of Fake News Publishers through Website Structure Data

Abstract

We propose a novel website structure based domain-level fake news detection model that has performance results surprisingly comparable to that of existing content-based methods. Through feature analysis, we highlight that fake news sites have more clustered subpages and more *ads* links, whereas traditional news sites are more substantive and more likely to contain *staff* links. We then illustrate that the structural model has a higher overall false positive rate compared to content-based methods, which have a higher false negative rate for domains that are more recent, more popular, and is conservative-leaning. Additionally, we also show that all model performance is dependent on the strictness in definitions of fake and traditional news sites. Specifically, model performance is higher when these definitions are more restrictive. Finally, we demonstrate that the performance of existing content-based models improve significantly by incorporating structural features, particularly when the definitions for fake and traditional news sites are lax.

1 Introduction

In the United States, many political pundits and media scholars alike have cautioned against the rising influence of fake news (Silverman, 2017; Balmas, 2014), stressing that the spread of false information weakens the legitimacy and public trust in the established political and media institutions. Outside of the U.S., fake news is also culpable of contributing to Brexit in Europe (Kucharski, 2016), the rising hate, violence, and nationalism in Indonesia (Kwok, 2017), and endangering the election integrity of nations in Europe and Latin America (Fletcher et al., 2018; Alimonti and Veridiana, 2018). Indeed, fake news, backed by armies of social bots, disseminates significantly faster and deeper than mainstream news (Shao et al., 2017). Additionally, subsequent research also suggests that it is difficult for the general public to tell fake news apart from credible content and that repeated exposure causes readers to perceive false content as more accurate (Balmas, 2014). Thus, timely, scalable, and high-performing fake news detection automatons become a vital component in combating fake news.

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Thus far, researchers have leveraged linguistic attributes, user network characteristics, news articles' temporal propagation patterns, and deep-learning methods to build effective models that separate fake news from traditional news content (Zhou and Zafarani, 2018). Some methods classify fake news at article-level (Horne and Adali, 2017; Riedel et al., 2017; Vosoughi, Roy, and Aral, 2018; Shu et al., 2017) whereas others at domain-level (Yadav et al., 2010; Zahedi, Abbasi, and Chen, 2015). In this paper, we propose a novel website structure based model that detects fake news at domain-level. We then compare and contrast our model with existing content-centric benchmark classifiers. Our paper makes the following contributions:

- We first introduce a new taxonomy of fake and traditional news definitions using various boundaries rules. We then use it to consolidate and assign existing fake and traditional news domains identified by varied contributors into conceptually distinct subsets. The taxonomy also enables us to assess the robustness of model performance w.r.t different definitions of fake and traditional news sites.
- Next, we describe a novel website structure based domain-level fake news detection model and show that its performance is surprisingly comparable to content-based predictors. Further, by examining feature weights, we observe that fake news domains appear to have a more clustered subpage network and more *ads* links. In comparison, mainstream news sites have more unique subpages thus are more substantive. Likewise, through error analysis, we show that content-based models have a higher false negative rate for domains that are younger, more popular, and is conservative-leaning. In contrast, our model has a lower false negative rate in these 3 dimensions, and a slightly higher overall false positive rate.
- Additionally, we also illustrate that classifier performance is dependent on the taxonomy we introduced. Generally, all models perform better when the definitions for fake and traditional news sites are more restrictive.
- Finally, we show that the performance of existing content-based models improve significantly when combined with structural features, especially when definitions for fake and mainstream news sites are lax.

2 Related Work

Both the academia and industry have been putting in substantial resources to study and understand the characteristics of fake news in the hope of neutralizing its influence (Tambuscio et al., 2015; Vosoughi, Roy, and Aral, 2018; Vargo, Guo, and Amazeen, 2018). This ongoing effort includes both manual and automated approaches for detecting fake news.

For instance, in the industry, leading social media and tech firms including Facebook and Twitter (Hunt, 2017; Heath, 2016) rolled out features to use third party independent fact-checkers to assess the factualness of news articles and to alert users of "disputed information". In the academia, scholars have proposed the following machine learning approaches for automated fake news detection: i) using content-based features (Horne and Adali, 2017; Riedel et al., 2017) such as psycholinguistic attributes and writing style; ii) leveraging network characteristics such as user-to-article linkages (Tacchini et al., 2017), or network propagation patterns (Kwon et al., 2013); or iii) a combination of the two (Volkova et al., 2017; Buntain and Golbeck, 2017). Recently, deep learning based methods have also been proposed (Roy et al., 2018; Wang, 2017). Some of these proposals have translated into various automated tools and platforms for real-life applications including i) Hoaxy (Shao et al., 2016), a platform that visualizes the spread of both factual and doubtable claims or statements on Twitter, ii) the NELA-toolkit (Horne et al., 2018), a tool that uses content-based feature to classify and detect fake news article, iii) BS Detector (2017), a web-based browser plugin that searches all the links on a given webpage for references to unreliable news domains aggregated by Zimdars et al. (2016).

These fake news detection models and platforms were trained and built using various data sources: i) the set of popular fake news articles aggregated and published by *Buzzfeed.com* on Facebook during 2016 U.S. election (Silverman, 2016), ii) LIAR (Wang, 2017) which includes 12.8K annotated short statements sampled from politifact.com, iii) CREDBANK (Mitra and Gilbert, 2015), a large scale crowdsourced dataset of approximately 60 million tweets with annotated credibility scores, iv) selected articles from *Factcheck.org* and *Snopes*, and finally, v) various fake news domains lists (Zimdars, 2016; Politifact Staff, 2018).

In this paper, we build on these valuable contributions. We first consolidate existing fake and traditional news domain lists, and propose a new taxonomy of fake news definitions. This way, instead of relying on only one definition/list of fake news, we are able to identify the robustness of fake news classification methods w.r.t. such choices. We then introduce a novel website structure based domain level fake news detection model. Finally, we compare and contrast our classifier with existing benchmarks.

3 Data

We use 3 types of data: i) lists of fake and traditional news sites, and ii) homepages and subpages of the news sites present in the aforementioned lists, iii) webpages of news articles collected using the URLs embedded in the tweets about the presidential candidates during the 2016 election.

Lists of Fake and Traditional News Sites: We use the 5 distinct fake news lists and 3 traditional news list from both the academia and the press collected by Bozarth et al (2019). Fake news lists are from Zimdars et al. (2016), Allcott, Gentzkow, and Yu (2018), PolitiFact (2018), DailyDot (2018), and MediaBiasFactCheck (2018). Traditional news lists include Guo and Vargo (2018), Alexa (2019), and MediaBiasFactCheck (2018). These lists contribute to 1884 aggregated fake news sites, 8238 traditional news sites, and a combined total of 10K news sites. We refer readers to the original paper for a detailed description of these lists.

Homepage and Subpages: For each website i in the lists from the previous section, we use *scrapy* (2016), a Python crawler library, to scrape the content of i's homepage. Using this process, we obtain 8006 homepages [1]. We filter out domains with little content [2] and/or are for sale (see details in Section 5.1). This results in 7627 homepages. Next, for each of these homepages, we then use BeautifulSoup (Richardson, 2007), an XML parser, to aggregate the set of links contained in the homepage and filter out the subset of links leading to external websites. Next, we recursively crawl the subpages associated with the remaining links and extract their content to aggregate additional links. We repeat this recursion 3 times (i.e. a depth of 3 with the homepage at depth=0). We collect a total of 2.1 million unique subpages.

News Article Webpages and Tweets: We obtain 174K unique articles shared by 711K Tweets about Clinton or Trump between December 31, 2015, and January 1, 2017 on Twitter from Bozarth et al. (2019). Each article has a record with: i) the shortened URL, ii) the original URL, iii) domain name (e.g. nytimes.com), iv) title of the document, v) body of the document, and (vi) the corresponding tweet object. All URLs match one of the domains in the aforementioned fake and traditional news sites lists.

4 Taxonomy of News Site Definitions and Gold Standard Label Space

Prior research has demonstrated that the attributes of supervised-learners (e.g. performance) are significantly dependent on the quality of gold standard labels (Lease, 2011). Within the context of fake news classification, the gold standard labels used by existing tools and classifiers differ. For instance, Horne et al. (2018) get fake news articles from Burfoot and Baldwin (2009), and *Buzzfeed*, they then filter out articles not from the fake news domains listed by Zimdars et al. (2016). Riedel et al (2017) use the dataset provided by fakenewschallege.org which originated from Craig Silverman's list of top fake news sites (2016). While both models predict at article-level, their data were restricted by fake news domain lists. Additionally, the BS Detector (2017) collects and detects fake news at domain-level.

[1] The remaining 2K domains either timed-out during scraping or returned 404 errors (Not Found), 502 (Bad Gateway), 503 (Service Unavailable), et cetera.

[2] We remove domains with < 1000 in character count of actual text; 178 or 2.2% domains are removed here.

Given these considerations, in this section, we first introduce a taxonomy of fake and traditional news definitions using various boundary rules (or, strictness in definition). We then consolidate and reassign domains aggregated by Bozarth et al. (2019), described in Section 3, into 7 distinct fake news lists and 2 traditional news lists. Finally, we provide additional insights into the characteristics of these lists.

Fake news lists with varied strictness in definition: Intuitively, if a domain is included by multiple fake news domain lists, then this domain has met the various criteria or definitions set by the corresponding list contributors. In other words, if a domain is recorded by multiple lists, there is a consensus that it is a fake news site. Similarly, fake news scholars employ mixed rules to divide websites into subcategories such as *fake*, *bias*, *junksci*, *hate*, et cetera (Zimdars, 2016). Thus, a domain that's assigned by multiple lists into the *fake* subcategory should be more decidedly fake than another that's assigned to *fake* fewer times. Finally, a subset of the domains present in these lists is clearly noted to have both questionable articles but also credible content (e.g. domains that belong to the *unreliable* subcategory in the list by Zimdars et al.). Given these observations, we use 3 types of rules to impose rigor in the definition of fake news sites. We define the following:

- Domains included by at least 1, 2, or 3 fake news lists.
  1. $LOOSE$ 2. $LST|2$ 3. $LST|3$
- Domains assigned by at least 1, 2, or 3 fake news lists into the *fake* subcategory.
  4. $FAKE|1$ 5. $FAKE|2$ 6. $FAKE|3$
- Domains suggested by at least 1 list to contain some false claims but also credible content.
  7. $MIXED$

The visualization of the 7 sets of fake news domains, generated using the aforementioned taxonomy, and their sizes are shown in Figure 2.

Traditional news lists with varied strictness in definition: Similarly, we also impose boundaries on what's considered a traditional news site and define the following:

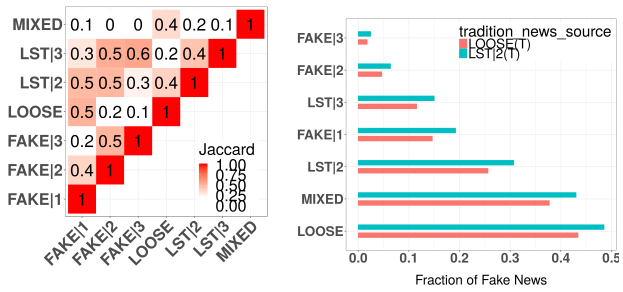
- Domains listed by at least 1, or 2 traditional news lists.
  1. $LOOSE(T)$ 2. $LST|2(T)$

The sizes for $LOOSE(T)$ and $LST|2(T)$ are 8238 and 1008.

For our paper, we compare and contrast our structure-based models against existing baseline models (e.g. performance and performance-stability) using each golden labelset pair (f,t) where $f \in \{MIXED, LOOSE, LST|2, LST|3, FAKE|1, FAKE|2, FAKE|3\}$, $t \in \{LOOSE(T), LST|2(T)\}$. Results obtained can give us more insights into how different classifiers behave when trained on fake and traditional news sites labels generated using varied definitions.

Additional Observations: Here, we provide 2 highlights that demonstrate the differences between generated lists: i) Jaccard similarity, and ii) prevalence.

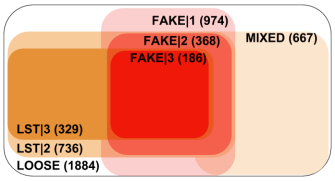
Jaccard similarity scores, defined based on the set of domain overlap, between each pair of fake news lists are shown in Figure 1a. We observe the maximum score is 0.6 (between



(a) Jaccard similarity between lists of fake news sites generated using varied definitions.

(b) Fake news prevalence using lists of fake and traditional news sites generated using varied definition.

Figure 2: Fake News with Varied Strictness in Definition.
For instance, $LST|2$ is the set of domains listed by at least 2 fake news lists and $FAKE|2$ is the set assigned by at least 2 lists into the *fake* subcategory.



$FAKE|3$ and $LST|3$). The median score is only 0.42, suggesting these lists are significantly different.

Given (f,t), we measure the prevalence of fake news using $\frac{|s_f|}{|s_f|+|s_t|}$, where $|s_f|$ and $|s_t|$ as the total number of tweets, or shares, containing URLs from domains in f and t. As shown in Figure 1b, depending on f and r, over 40% (when using the most lax definition for fake news sites) or less than 3% (when using the most restrictive definition) of all news articles shared on are from fake news sites. Conceptually, prevalence generated using $(LOOSE, LST|2(T))$ and $(FAKE|3, LOOSE(T))$ can serve as upper and lower bounds on what you would find using existing raw lists. Additionally, we observe domains in $MIXED$ contribute to a large fraction of tweet shares, suggesting that low-quality news content (though, not necessarily fake), were shared frequently during the 2016 election.

5 Models

In this section, we first introduce 3 novel `structure` models: i) $s1|basic$, ii) $s2|subpage$ and iii) $s3|complete$. We then describe existing `baseline` fake news classifiers proposed by other papers including: iv) *NELA* (Horne et al., 2018), and v) *RDEL* (Riedel et al., 2017). Finally, we update baseline predictors by including structural features from $s3|complete$; we refer to the modified `baseline+` models as vii) $NELA+$, viii) $REDL+$, and ix) *voting*.

5.1 Structure-based Model Using Homepage

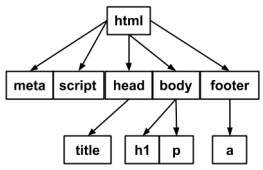
In this subsection, we outline all the features of $s1|basic$ generated using a website's homepage: i) URL characteristics,

ii) homepage auxiliary data, and iii) homepage HTML elements and paths.

URL Characteristics: For each website i, we generate the following features: $https_i = \{1,0\}$, $suffix\_count_i$, $domain\_length_i$, and $top\_level_i = \{1,0\}$ where $top\_level \in \{com, org, net, edu, gov, other\}$. To elaborate, a website's URL typically contains a protocol which is generally either "https" or "http"; the former is a secure protocol the later unprotected. A news site that uses "https" suggests better privacy and website validity, thus we include $https_i$. Next, a URL also has a top-level domain, and to ensure better high-level website categorization, certain top-level domains are restricted to specific entities (e.g.".gov" is restricted to government websites and ".edu" to established educational institutions), therefore we incorporate top_level_i. Further, prior research on phishing (Alkhozae and Batarfi, 2011) demonstrates that malicious websites often attempt to fool people by using domains closely resembling those of well-known sites. For fake news sites, we see that "abcnews.com.co", an identified fake news site, is a clear attempt to confuse users with the legitimate "abcnews.com". The former has an additional domain suffix compared to the later which can be differentiated by $suffix\_count_i$. Finally, easy-to-remember and short domain names are increasingly harder to find and more expansive to own (Pinsky, 2017) so we also include $domain\_length_i$.

```
<!DOCTYPE html>
<html>
  <meta charset="UTF-8"/>
  <script>...</script>
  <head>
  <title>HomePage Title</title>
  </head>
  <body>
    <h1>First Heading</h1>
    <p>First paragraph.</p>
  </body>
  <footer>
  <a href='/job'>Jobs</a>
  </footer>
</html>
```

(a) Basic HTML representation of a webpage.



(b) DOM tree representation of the webpage in Figure 3a. An example tagpath is $html \rightarrow body \rightarrow h1$

Homepage Auxiliary Data: We presume that mainstream news organizations possess more financial resources, thus their websites are likely better designed and managed compared to fake news sites. Webpages are represented by underlying HTML objects (see Figure 3a), and a well-designed webpage can manifest in the form of having more responsive front-end scripts and elaborate style sheets in the HTML object. Additionally, mainstream news organizations can include certain links in their homepage that may be less common in fake news sites (e.g. job opportunities). Here, we divide these auxiliary features into 2 subtypes: i) homepage style and scripts, and ii) homepage link categories.

Homepage Style and Scripts: For a given website i and tag s, where $s \in S$ and $S = \{meta, script, style, noscript, embed, params, object\}$ is a set of common HTML tags used for programming scripts, style, and metadata. We write s_i^{cnt} and s_i^{len} as the total number of s in i's homepage and the aggregated character count of the inner HTML text in s.

Homepage Link Categories:. Mainstream news sites are likely to include routine links such as privacy policies, contact form, about us page, career opportunities, donation information, and subscription links in their homepage. In comparison, a prior study by Starbird (2017) suggests that many fake news sites were created in order to "sell overpriced vitamins" to its viewers, thus fake news sites are likely to have more ads or an online store. Here, let $C^1 = \{privacy, ads, contact, about, career, donation, subscription, store\}$, and $c^1 \in C^1$. for each domain i, we set $c_i^1 = 1$ if there is a matching clickable short link[3] in i's homepage[4]. As an example, the variable $career_i = 1$ indicates that i's homepage contains a link that leads to job postings. Similarly, we also derive the top 100 most common 1,2-grams[5] using only text in short links of all domain homepages, filtering out the tokens in C^1. We denote this list as C^2. Let $c^2 \in C^2$, we set $c_i^2 = 1$ if there is matching short link in i's homepage.

Homepage HTML Tags and Path A DOM tree (as shown in Figure 3b) is a tree structure model that represents an HTML document (i.e the underlying layout of a webpage). In this subsection, we define features associated with DOM tree elements and tree paths.

Element Tags: HTML documents generally contain a head, body and footer section. Each section then contains additional nested web elements, or tags. Common web tags can be broadly divided into seven categories: i) text and font, ii) images, iii) frames, iv) form elements, v) links or navigation, vi) lists, and vii) multimedia (W3schools Contributors, 2019). Here, for $s^1 \in \{header, body, footer\}$, and $s^2 \in \{text, image, frames, form, links, lists, multimedia\}$, we denote $cnt(s^1, s^2)$ as the total number of tags from category s^2 that exist in section s^1. Likewise, $depth(s^1, s^2)$ is the max tree depth of tags from category s^2 that exist in section s^1.

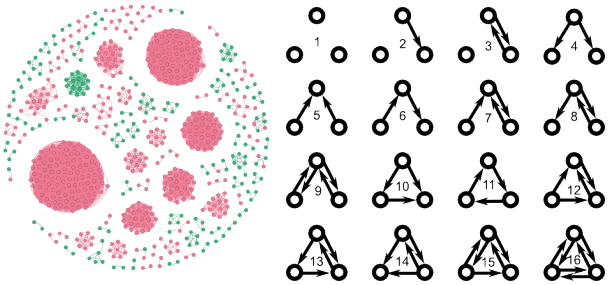
Tags Path: DOM tree path features have been shown to be effective in detecting phishing websites (Joshi et al., 2003). More specifically, impostor-websites are known to mimic or replicate the HTML template of an authentic website in order to trick users. Within the context of fake news, we observe certain fake news domains have the exact same homepage layout (e.g. vaccines.news and mediafactwatch.com). Here, given a domain i's homepage DOM tree, we determine all the paths from the root (i.e. the $< html >$ tag) of the tree to each leaf node as denote it as $P_i = \{p_0, p_1 ...\}$. Referencing Figure 3b, an example tagpath is $html \rightarrow body \rightarrow h1$. For P_i, We calculate i) total tagpaths in P_i, and ii) total unique tagpaths. Further, given $|P_i| = \{|p_0|, |p_1|..\}$ where $|p_i|$ is the

[3]Here, we only use links that have fewer than 20 character count in the displayed text to avoid news article links (i.e. link <a>politicsis included, but <a>super long article title nameis excluded).

[4]We use the following keywords for matching: about (about), privacy (privacy), donate (donat, "support us"), contact (contact), career ("work with", "work for", "join our team", "for us", "job"), staff ("staff", "our team", "contributor"), store ("store", "shop", "buy ", "product"), and ads ("adverti", " ad"). We obtain these keywords by manually searching through the most common 1,2-grams.

[5]For instance, "politics" is one of the common terms appeared in a lot of links (e.g. https://nytimes.com/politics/)

Figure 4: Note, Figure 4a): Traditional news domains are in red, fake in green. An edge exists between 2 domains if their tagpath cosine similarity >= 0.9. Figure 4b): three node motifs.



(a) (b)

length of tagpath p_i, we also calculate iii) average and iv) max length of $|P_i|$, v) gini coefficient (Yitzhaki, 1979), vi) skewedness and vii) kurtosis of $|P_i|$. Additionally, by treating each tagpath as a text token, we use the bag-of-words approach used in text analysis (Wallach, 2006) to calculate cosine similarities, $cos_{i,j}$, between each pair of P_i and P_j where domain $i \neq j$. We observe that 2923 domains have $cos >= 0.9$ with at least 1 other domain. We manually check groups of these domains and remove 201 that are onsale[6].

Finally, to gain additional insights into using tagpaths as features, we also use Gephi (Bastian, Heymann, and Jacomy, 2009) and the Fruchterman Reingold layout (1991) to plot the graph generated using domains from *LOOSE* and $LST2|(T)$ that also have $cos >= 0.9$ with at least 1 other domain (from the same 2 lists) in Figure 4a. Domains from *LOOSE* are colored in green and $LST|2(T)$ in red. As shown, we see several large clusters of traditional news domains and smaller clusters of fake news domains. This suggests that certain subsets of traditional news sites (e.g. columbiatribune.com and morningsun.net) share the same homepage template. Same for fake news sites.

5.2 Structure-based Model Using Subpages

In this subsection, we describe all features of $s2|subpage$ generated using a website's subpage-to-subpage linkages. To elaborate, each website has a homepage and many additional subpages. By randomly sampling 10 fake news sites, we observe that the majority of them have limited navigation depth which differs from well-known news sites. We postulate that differences in the subpages networks between mainstream and fake news sites can be leveraged to differentiate them. Here, we assign features into 2 subcategories: i) characteristics of the entire network, and ii) motifs.

We use the following process to generate a subpage-to-subpage network. For each website i, we define a directed graph $G_i = \{V_i, E_i\}$ where V is the entire set of crawled webpages. Let $u \in V$, $v \in V$, and $u \neq v$, a directed edge $e_{u,v}$ exists in E if there is a clickable link from the subpage u to v.

Network Characteristics: Given i, we determine i) the size of the network ($|V_i|$), ii) the average number of unique links per subpage ($\frac{|E_i|}{|V_i|}$), iii) the local clustering coefficient, iv) betweenness centrality, normalized and non-normalized Gini coefficients (Yitzhaki, 1979) for the entire v) indegree and vi) outdegree distributions, vii) number of strongly connected components, viii) the size of the largest strongly connected component, and ix) number of communities in G_i using the Louvian (De Meo et al., 2011) method. These are common metrics used to analyze social networks. Their definitions, operationalization, and applications can be found in work by Wasserman, and Easley&Kleinberg (1994; 2010).

Motifs: Network motifs, as shown in Figure 4b, are small subgraphs that frequently occur in graphs (Milo et al., 2002). That is, motifs are like lego pieces of large complex networks. Leskovec et al. (2006) demonstrate that the frequency distribution of motifs differ for networks of varied categories (e.g. book vs. music recommendation networks). Here, we extract the number of motifs for G_i using *snap* (Leskovec and Sosič, 2016), a Stanford network analysis library. Given a node size of three, there exist 16 unique motifs as shown in Figure 4b. For G_i, we calculate the total number of motifs and write them as $M_i = \{m_{i,1}, m_{i,2}, ...\}$ where $m_{i,1}$ is the total number of motif type 1 (as shown in Figure 4b) normalized by the number of edges $|E|$.

5.3 Structure-based Model Complete

We aggregate all features in $s1|basic$ and $s2|subpage$ into a single model $s3|complete$; it contains features generated using a domain's homepage (i.e. URL characteristics, homepage auxiliary data, and homepage HTML elements and paths) as well as subpage-to-subpage linkages (network characteristics and motifs).

5.4 Baseline and Baseline+ Models

We first describe 2 existing baseline models, *NELA* and *RDEL*, that use content-based features to classify fake and mainstream news [7]. We then describe `baseline+` models that combine `baseline` models with structural features.

News Landscape (NELA) Toolkit: This classifier is provided by Horne et al. (2018), and we denote it as *NELA*. It uses the following 3 distinct dimensions of text-based features to predict false news content: i) style features including punctuation (e.g. exclamation marks), verb tense, pronoun usages, et cetera, ii) psycho-linguistic features such as sentiment scores using LWIC (Pennebaker, Francis, and Booth, 2001), SentiStrength (Thelwall, 2017), et cetera, and iii) content complexity features including readability (Mc Laughlin, 1969), dictionary size, average word length. We refer readers to the original paper for the complete list of 100+ features. We note that Horne et al. use Linear Support Vector Machine (SVM) (Suykens and Vandewalle, 1999) and Random Forest (Liaw, Wiener, and others, 2002) as their classification algorithms.

[6]These domains which share the same template and are being sold by www.networksolutions.com,www.mydomaincontact.com, et cetera

[7]We also contacted Volkova et al. (2017) to request their model's code repository.

Ngram-based Model: This model is proposed by Ridel et al. (2017) and we write it as *RDEL*. The authors first tokenize text from news articles and extract the most frequent vocabularies [8]. Then, for each news article, they construct the corresponding term frequency-inverse document frequency (TF-IDF) (Ramos and others, 2003) vectors for article title and body separately, and compute the cosine similarity between the 2 vectors. Finally, the authors use Multi-layer Perceptron, implemented in *scikit-learn* (Pedregosa et al., 2011), to classify fake and real news articles.

Baseline+ Models: We append all structure features from $s3|complete$ to the base *NELA* and *RDEL* models, and denote the updated models *NELA*+ and *RDEL*+. Further, we also denote *voting* classifier which takes prediction results from $s3|complete$, *NELA* and *RDEL*, and then outputs the majority vote. These 3 are our `baseline+` models.

6 Classification Performance, Feature Weights, and Error Analysis

6.1 Data and gold standard labelsets

Referencing Section 4, we use the following process to generate subsets of data corresponding to each gold standard labelset pair (f,t) [9]. For `structure` models, we denote $\mathscr{D}^{s}_{(f,t)}$ as the subset of data that contains structural information and features of a domain i only if $i \in f \vee i \in t$. Similarly, for `baseline` and `baseline+` models, we denote $\mathscr{D}^{b}_{(f,t)}$ as the subset of data that contains news articles and features of domain i only if $(i \in f \vee i \in t) \wedge i \in D$ where D is the complete list of +7K domains that have existing homepages(see Section 3). This is to ensure performance comparison between `structure` and `baseline` models are done using only the domains with homepages [10].

6.2 Performance Evaluation and Comparison

Evaluation metrics: Given the class imbalance (e.g. mainstream to fake news ratio of 4.5:1 for LOOSE and LOOSE(T)), we use ROC AUC (Bradley, 1997; Kotsiantis et al., 2006) to evaluate classifier performance.

Prediction Adjustment: Classification of `baseline` and `baseline+` models is at the article-level (i.e. given an news article, is it fake?), whereas that of `structure` is at domain-level (i.e given a domain, is it a fake news site?). We use majority-voting of article predictions to adjust prediction results such that all classification is done at domain-level. As a robustness check, we compare performance results for `baseline` models that predict at i) article-level and at ii) domain-level using the adjusted majority-voting approach. Results suggest that AUC scores between them are comparable with domain-level prediction actually performing slightly better (e.g. on average, *NELA* using majority-voting has a 0.03 higher AUC score).

Performance Comparison: We tried several different supervised-learning algorithms [11] with parameter tuning [12]. Our results show that the Random Forest classifier outperforms the other algorithms for all `structure` models. Thus, here we report only on results generated by Random Forest for `structure`. We also report on the best results for *NELA* (Random Forest) and *RDEL* (Multi-layer Perceptron). AUC scores for all models are summarized in Figures 5a and 5b. The y-axis denotes the AUC scores, the x-axis lists F, and finally, the sub-headers are the T values.

Figure 5a contains results for `structure` and `baseline` models. Structure-based models are colored in purple, `NELA` models in red, and `RIDL` in green. As shown, all models show higher performance when trained using more restrictive definitions of fake and real news sites (e.g. models trained using $(FAKE|3, LST|2(T))$ have considerably higher AUC measures than those trained on $(MIXED, LOOSE(T))$). This pattern is, however, more mild in `structure` models, suggesting that our models are more robust to different news site definitions. Further, $s3|complete$ significantly outperforms the original *NELA* model when the boundaries conditions for fake news sites are less restrictive, but the *NELA* model does better with more strict boundaries. Further, the *RDEL* model outperforms the other classifiers when the $t = LOOSE(T)$, but is comparable to $s3|complete$ when $t = LST|2(T)$. Finally, we note that performance of *NELA* and *REDL* here are also comparable to what's described in the original papers (Horne and Adali, 2017; Riedel et al., 2017).

Next, in Figure 5b, we compare `baseline` models (drawn in solid lines) to `baseline+` models (dashed lines). We observe that *NELA*+ outperforms base *NELA* model when definitions are lax (e.g. when $f = LOOSE$ or $f = FAKE|1$). Further, *RDEL*+ outperforms *RDEL* across all pairs of (f,t). Finally the *voting* classifier outperforms all classifiers except for *RDEL*+.

These observations suggest that i) classifier performance is dependent on the taxonomy of fake news definitions;

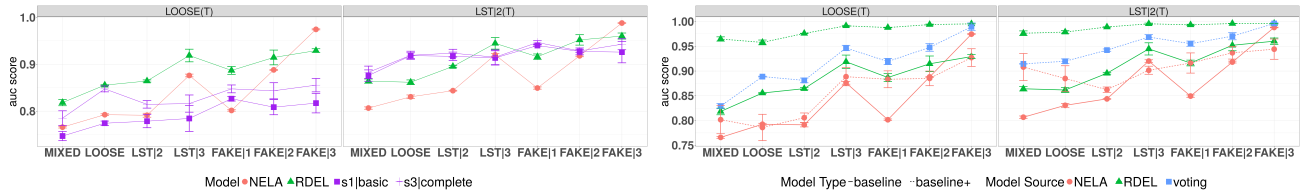
[8] Tokens of newspapers names are removed (e.g. *daili beast*).

[9] Here, if a news site exists in both lists f and t, it's treated as a fake news site. In other words, f precedes t.

[10] We also trained `baseline` models using all domains. Performance is comparable, thus results are omitted for brevity.

[11] We select the following commonly used supervised machine learning algorithms: i) Linear Regression (Press and Wilson, 1978), ii) Linear Support Vector Machine (SVM) (Suykens and Vandewalle, 1999), iii) Random Forest (Liaw, Wiener, and others, 2002), and iv) Multi-layer Perceptron (Gardner and Dorling, 1998), all of which are implemented by Python's *Scikit-learn* library (Pedregosa et al., 2011). Logistic Regression was used by Tacchini et al. (2017) in fake news classification, Linear SVM and Random Forest by Horne et al. (2018). Additionally, Multi-layer Perceptron is a shallow artificial neural network used by Riedel et al. (2017) in the Fake News Challenge stance detection task.

[12] For parameter tuning, we combine each algorithm with SearchGridCV (Pedregosa et al., 2011), a function that searches over specified parameter space for the given estimator. Here, we use i) StratifiedKFold ($cv = 5$) as the cross-validation generator and ii) Area Under the Receiver Operating Characteristic Curve (ROC AUC), described in Section 6.2, as our scoring metric in SearchGridCV. For each dataset, $\mathscr{D}_{(f,t)}$, We split it 80/20 for training and validation. We run SearchGridCV using the 80-split dataset (balanced using upsampling which is shown by prior work (Kotsiantis et al., 2006) to improve performance), and then test the best estimator returned on the remaining 20-split.



(a) ROC AUC comparison between `structure` and `baseline`. (b) ROC AUC comparison between `baseline` and `baseline+`.

ii) structure based methods' performance is surprisingly comparable to content-based methods; and iii) the performance of existing content-based models improve significantly when combined with structural features, especially when definitions for fake and mainstream news sites are lax.

6.3 Feature Weights

In this section, we first assess the volatility of feature weighting of models trained using different golden standard labels. For each classifier c, let $W^c(f1,t1)$, and $W^c(f2,t2)$ denote the feature weights of c trained using the labelset $(f1,t1)$ and $(f2,t2)$, we use Pearson's correlation coefficient (Lawrence and Lin, 1989) to evaluate the correlation between $W(f1,t1)$, and $W(f2,t2)$. We observe that the median correlation scores are 0.9 and 0.48 for *NELA* and *NELA+* respectively, suggesting that features weights are more stable across pairs of (f,t) for the original model. A manual inspection shows that structure features have more significant weights (both negative and positive) when *NELA+* is trained using (f,t) pairs of more lax definitions. In comparison, the median correlation score is 0.61 for $s3|complete$, suggesting that feature weights for structural models are more volatile across pairs of (f,t).

Next, we extract the most positive and negative features of each *structure* model [13]. We first select the top 25 [14] most positive or negative features for each (f,t). We then determine the top 5 most frequently occurring features across all pairs of (f,t) and write them as *stable* features. Similarly, we also select a subset of least frequently occurring features and denote them as *volatile* features. Both *stable* and *volatile* features for each *structure* model are listed on Table 1. As shown, fake news domains appear to have higher subpage-to-subpage clustering: i) average clustering coefficient and ii) motif 16 are both stable positive features. Additionally, fake news sites also have more skewed tagpath depth and *ads* links. In comparison, traditional news sites are associated with having more unique subpages. This is consistent with our hypothesis that news organizations are better staffed and therefore have a more substantive website.

6.4 Error Analysis

In this section, we assess whether models perform better or worse on classifying certain domains by correlating prediction errors with a domain's i) ideological-leaning, ii) popularity, and iii) age. We obtain data on domain age, popularity, and ideological-leaning from work by Bozarth et al. (2019). Here, given $c \in \{NELA, RDEL, s3\text{-}complete\}$ (i.e. we compare our complete structure classifier to models from existing papers), fake and traditional news site lists pair (f,t), and the corresponding dataset $\mathscr{D}_{(f,t)}$, we derive $E^{fp}(f,t,c)$ as the set of domains in $\mathscr{D}_{(f,t)}$'s validation fold that has been classified incorrectly by c as *false positives*. Similarly, we denote $E^{fn}(f,t,c)$ for *false negatives*.

Age: Given (f,t), we first partition each domain i where $i \in t \vee i \in t$ into 4 bins using age: i) age unknown, ii) 33% percentile (i.e. age percentile calculated to be between 0 and 33%), iii) 66% percentile, and iv) 100% percentile (domains at 100% percentile are the oldest). Next, let $bin^n_{f,t}$ be the set of domains in bin n. Then, for each classifier c, we calculate its fraction of false positive error for bin n using $\frac{|bin^n_{f,t} \cap E^{fp}(f,t,c)|}{|bin^n_{f,t}|}$ where $|bin^n_{f,t} \cap E^{fp}(f,t,c)|$ is the number of domains in $bin^n_{f,t}$ that are also in $E^{fp}(f,t,c)$. We then calculate the fraction of false negative error using $E^{fn}(f,t,c)$. We repeat the process for all combinations of (f,t) and c.

Popularity: Similar to age, we assign errors into 4 bins using popularity: i) 25% percentile, i) 50% percentile, i) 75% percentile, and i) 100% percentile. We then calculate the fraction of error for each bin for each combination of (f,t) and C using the process described above.

Bias (ideological-leaning): Likewise, we divide errors into bins: i) unknown, ii) conservative iii) liberal, and iv) center. We then repeat the aforementioned procedure.

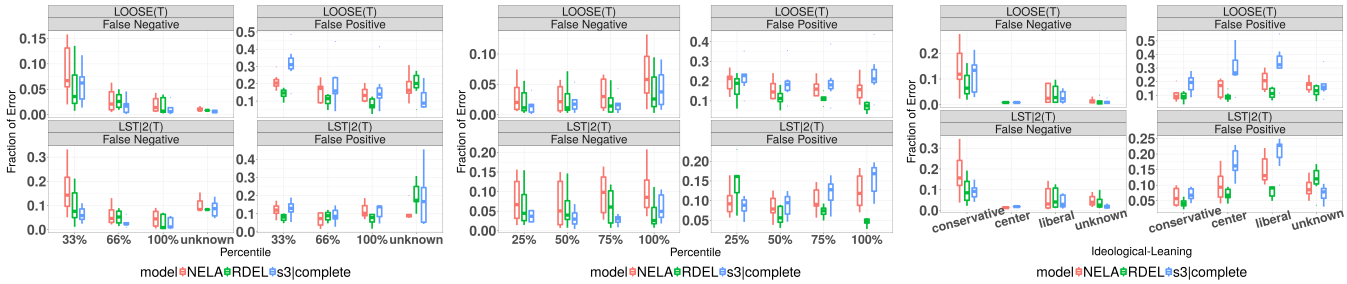
Error analysis results for *age*, *popularity*, and *bias* are shown in Figures 6a, 6b, and 6c. The x-axis denotes the bins, the y-axis fraction of error for each bin; each color corresponds to a unique model. As shown, the $s3|complete$ model has an overall lower false negative rate and a higher false positive rate, especially when $t = LOOSE(T)$. For instance, the average false positive error rate for unpopular domains (less than $< 33\%$ percentile by popularity) by $s3|complete$ is over 20% when $t = LOOSE(T)$. We also observe a high false positive rate for liberal-leaning domains. Given Bozarth et al. (2019) have shown that unpopular domains are less likely to be included in fake news lists and conservative critics' claim of liberal bias in the academia, it's possible that these domains are actually fake but not yet labeled. In comparison, the original *NELA* model has a significantly higher false negative rate for more recent domains, more popular

[13] Here, we use results from the Logistic Regression classifier given Random Forest does not supports signed feature weights.

[14] We select only the top 10 features instead for $s2|subpage$ given it has fewer than 50 features.

Table 1: Top Positive and Negative Features for each *structure* model. Note, the terms "body", "head", "iframe", "footer" refer to HTML elements; images of motifs are located in Figure 4b.

model	type	most.positive	most.negative
basic	stable	ads; number of links; press; onlin; report	number of navigation links; domain suffix; newslett; polit; local news
	volatile	number of nontext tags in the body; number of stlye sheets; editori; food; term; facebook	photo; depth of head element; tv; number of list tags in body and footer; donate
subpage	stable	number of unique outgoing links; average number of / in links; average clustering coefficient; gini for out degree distribution; motif 10	gini for indegree distribution; number of unique subpages crawled; motif 7 and 13
	volatile	motif 16	motif 4;
complete	stable	number fo links; average number of / in links; tagpath depth distribution skewness; ads; average clustering coefficient	number of unique subpages crawled; tagpath depth standard deviation; domain suffix; local news; sales
	volatile	max tagpath depth; copyright; food; number of frames (e.g. iframe)	number of image tags in head; staff; blog; motif 16; number of object tags;



(a) Error analysis using age.

(b) Error analysis using popularity.

(c) Error analysis using ideology

Figure 6: Classification error analysis using a domain's i) age ii) popularity, and iii) ideological-leaning.

domains, and domains with conservative-bias. Further, error attributes for *RDEL* is comparable to *NELA*, suggesting that models using same feature types (i.e. content) share similar error characteristics.

7 Discussion

In this paper, we introduced a new taxonomy of fake and traditional news definitions using varied boundary conditions. We then proposed a novel website structure based domain-level fake news detection model that had demonstrated surprisingly comparable performance to content based benchmarks. This highlights the potential of determining quality without inspecting the content and through simpler methods with ideology-agnostic features. Alternatively, this finding suggests content based methods have a long way to go if simple structural features provide more predictive power. Yet, it is also worthwhile to note that fake news producers can, in the future, invest more resources as they gain in popularity and revenue to improve their websites.

We then highlighted differences in structural features between fake and traditional news domains. We found that fake news sites have more clustered subpages network (i.e. each subpage is linked to all the other subpages), more *ads* links, more iframes. In contrast, mainstream news domains are associated with more unique subpages, and *staff* links. Moreover, certain subsets of fake (or traditional) news sites share the same exact HTML template. Additionally, through error analysis, we showed that models generated using a particular category of data (e.g. content) shared similar errors. More specifically, content-based models had a higher false negative rate for younger, more popular, and conservative-leaning domains. In contrast, structure based model had a slightly higher overall false positive rate. Finally, we emphasize that performance of existing content-based models improved significantly when combined with structural features, especially when definitions for fake and mainstream news sites were lax. In other words, structural factors can be combined with content features to better detect news sites that provide articles that are not necessarily completely false but nevertheless low in content quality.

There are several limitations to our work. First, while content-based methods are the most prevalent, high-performing models that use other categories of features such as user network also exist. Future work should include the comparison of those models and our structure based approach. Additionally, focusing on classification errors, it's apparent that existing models are biased, in different ways, against domains with varied ideological-leanings. This is related to the ongoing conversation on how black-boxed, unaccountable machine learning models lead to increased inequality. Considering the critical importance of ideology in political communications, future work on fake news detection should assess model biases particularly in this dimension and make the results available. Finally, it's also possible some of the false positive domains we found are actually unlabeled fake news sites. Future work should also include collaborating with fake news list aggregators to identify additional fake news domains.

References

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

opponents.

Alkhozae, M. G., and Batarfi, O. A. 2011. Phishing websites detection based on phishing characteristics in the webpage source code. *International Journal of Information and Communication Technology Research* 1(6).

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.

Bastian, M.; Heymann, S.; and Jacomy, M. 2009. Gephi: an open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media*.

Bozarth, Lia and Saraf, Aparjita and Budak, Ceren. 2019. Higher ground? how groundtruth labeling impacts our understanding of the spread of fake news during the 2016 election. [Online; accessed 4-May-2019].

Bradley, A. P. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition* 30(7):1145–1159.

Buntain, C., and Golbeck, J. 2017. Automatically identifying fake news in popular twitter threads. In *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, 208–215. IEEE.

Burfoot, C., and Baldwin, T. 2009. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, 161–164.

De Meo, P.; Ferrara, E.; Fiumara, G.; and Provetti, A. 2011. Generalized louvain method for community detection in large networks. In *2011 11th International Conference on Intelligent Systems Design and Applications*, 88–93. IEEE.

Easley, D.; Kleinberg, J.; et al. 2010. *Networks, crowds, and markets*, volume 8. Cambridge university press Cambridge.

Fletcher, R.; Cornia, A.; Graves, L.; and Nielsen, R. K. 2018. Measuring the reach of fake news and online disinformation in europe. *Reuters Institute Factsheet*.

Fruchterman, T. M., and Reingold, E. M. 1991. Graph drawing by force-directed placement. *Software: Practice and experience* 21(11):1129–1164.

Gardner, M. W., and Dorling, S. 1998. Artificial neural networks (the multilayer perceptron)a review of applications in the atmospheric sciences. *Atmospheric environment* 32(14-15):2627–2636.

Gitlab Contributors. 2017. bs-detector.

Heath, A. 2016. Facebook is going to use snopes and other fact-checkers to combat and buryfake news. *Business Insider*.

Horne, B. D., and Adali, S. 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.

Horne, B. D.; Dron, W.; Khedr, S.; and Adali, S. 2018. Assessing the News Landscape: A Multi-Module Toolkit for Evaluating the Credibility of News. 235–238.

Hunt, E. 2017. Disputed by multiple fact-checkers: Facebook rolls out new alert to combat fake news. *The Guardian* 21.

Joshi, S.; Agrawal, N.; Krishnapuram, R.; and Negi, S. 2003. A bag of paths model for measuring structural similarity in web documents. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 577–582. ACM.

Kotsiantis, S.; Kanellopoulos, D.; Pintelas, P.; et al. 2006. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering* 30(1):25–36.

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

Kwok, Y. 2017. Where memes could kill: Indonesias worsening problem of fake news. *Time, January* 6.

Kwon, S.; Cha, M.; Jung, K.; Chen, W.; and Wang, Y. 2013. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th International Conference on Data Mining*, 1103–1108. IEEE.

Lawrence, I., and Lin, K. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 255–268.

Lease, M. 2011. On quality control and machine learning in crowdsourcing. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Leskovec, J., and Sosič, R. 2016. Snap: A general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology (TIST)* 8(1):1.

Leskovec, J.; Singh, A.; and Kleinberg, J. 2006. Patterns of influence in a recommendation network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 380–389. Springer.

Liaw, A.; Wiener, M.; et al. 2002. Classification and regression by randomforest. *R news* 2(3):18–22.

Mc Laughlin, G. H. 1969. Smog grading-a new readability formula. *Journal of reading* 12(8):639–646.

Milo, R.; Shen-Orr, S.; Itzkovitz, S.; Kashtan, N.; Chklovskii, D.; and Alon, U. 2002. Network motifs: simple building blocks of complex networks. *Science* 298(5594):824–827.

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

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71(2001):2001.

Pinsky, D. 2017. 8 smart tips for choosing a winning domain name.

Politifact Staff. 2018. Politifact's guide to fake news websites and what they peddle.

Press, S. J., and Wilson, S. 1978. Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association* 73(364):699–705.

Ramos, J., et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, 133–142. Piscataway, NJ.

Richardson, L. 2007. Beautiful soup documentation. *April*.

Riedel, B.; Augenstein, I.; Spithourakis, G. P.; and Riedel, S. 2017. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *arXiv:1707.03264 [cs]*. arXiv: 1707.03264.

Roy, A.; Basak, K.; Ekbal, A.; and Bhattacharyya, P. 2018. A deep ensemble framework for fake news detection and classification. *arXiv preprint arXiv:1811.04670*.

Scrapy, A. 2016. Fast and powerful scraping and web crawling framework. *Scrapy. org. Np*.

Shao, C.; Ciampaglia, G. L.; Flammini, A.; and Menczer, F. 2016. Hoaxy: A Platform for Tracking Online Misinformation. *WWW '16 Companion* 745–750. arXiv: 1603.01511.

Shao, C.; Ciampaglia, G. L.; Varol, O.; Flammini, A.; and Menczer, F. 2017. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592* 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.

Silverman, C. 2016. Here are 50 of the biggest fake news hits on facebook from 2016. *BuzzFeed, https://www. buzzfeed. com/craigsilverman/top-fake-news-of-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.

Suykens, J. A., and Vandewalle, J. 1999. Least squares support vector machine classifiers. *Neural processing letters* 9(3):293–300.

Tacchini, E.; Ballarin, G.; Della Vedova, M. L.; Moret, S.; and de Alfaro, L. 2017. Some like it : Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*.

Tambuscio, M.; Ruffo, G.; Flammini, A.; and Menczer, F. 2015. Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. In *Proceedings of the 24th international conference on World Wide Web*, 977–982. ACM.

Thelwall, M. 2017. The heart and soul of the web? sentiment strength detection in the social web with sentistrength. In *Cyberemotions*. Springer. 119–134.

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: A big data analysis of the online media landscape from 2014 to 2016. *new media & society* 20(5):2028–2049.

Volkova, S.; Shaffer, K.; Jang, J. Y.; and Hodas, N. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 647–653.

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

W3schools Contributors. 2019. HTML Tags Ordered by Category.

Wallach, H. M. 2006. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, 977–984. ACM.

Wang, W. Y. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. *arXiv:1705.00648 [cs]*. arXiv: 1705.00648.

Wasserman, S., and Faust, K. 1994. *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

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].

Yadav, S.; Reddy, A. K. K.; Reddy, A.; and Ranjan, S. 2010. Detecting algorithmically generated malicious domain names. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, 48–61. ACM.

Yitzhaki, S. 1979. Relative deprivation and the gini coefficient. *The quarterly journal of economics* 321–324.

Zahedi, F. M.; Abbasi, A.; and Chen, Y. 2015. Fake-website detection tools: Identifying elements that promote individuals' use and enhance their performance. *Journal of the Association for Information Systems* 16(6):448.

Zhou, X., and Zafarani, R. 2018. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*.

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