Identifying Disinformation Websites Using Infrastructure Features

Austin Hounsel
Princeton University
Jordan Holland
Princeton University
Ben Kaiser
Princeton University
Kevin Borgolte
Princeton University
Nick Feamster
University of Chicago
Jonathan Mayer
Princeton University

Abstract

Platforms have struggled to keep pace with the spread of disinformation. Current responses like user reports, manual analysis, and third-party fact checking are slow and difficult to scale, and as a result, disinformation can spread unchecked for some time after being created. Automation is essential for enabling platforms to respond rapidly to disinformation.

In this work, we explore a new direction for automated detection of disinformation websites: infrastructure features. Our hypothesis is that while disinformation websites may be perceptually similar to authentic news websites, there may also be significant non-perceptual differences in the domain registrations, TLS/SSL certificates, and web hosting configurations. Infrastructure features are particularly valuable for detecting disinformation websites because they are available before content goes live and reaches readers, enabling early detection.

We demonstrate the feasibility of our approach on a large corpus of labeled website snapshots. We also present results from a preliminary real-time deployment, successfully discovering disinformation websites while highlighting unexplored challenges for automated disinformation detection.

1 Introduction

In recent years, the Internet has made disinformation cheaper, easier, and more effective than ever before [37]. The same technologies that have democratized online content creation, distribution, and targeting are increasingly being weaponized to mislead and deceive. Russia deployed disinformation to interfere in the 2016 U.S. presidential election [14, 31, 45, 59, 67, 76], and political disinformation campaigns have also struck dozens of other nations [7, 8]. Other disinformation campaigns have been economically motivated, driving page views for advertising revenue, pushing products, or undermining competitors [43].

The major online platforms have not kept pace with the spread of disinformation. Responses to disinformation mostly rely on user reports, manual analysis, and fact checking, which are slow and difficult to scale [5]. Similarly, previous work on automated detection of disinformation has focused on textual and social graph features, which often rely on disinformation already being shared [1, 13, 25, 38, 41, 50, 73, 74, 78]. These responses give disinformation an asymmetric advantage, enabling it to spread and affect perceptions in the hours and days after it is first distributed—a critical period during which disinformation may be most effective [80].

In this paper, we explore the use of infrastructure features to detect disinformation websites. Our hypothesis is that while disinformation websites may be perceptually similar to authentic news websites, there may be significant non-perceptual differences in the domain registrations, TLS/SSL certificates, and web hosting configurations. We are motivated by prior work in the information security field that demonstrated viable early detection for malware, phishing, and scams using machine learning and a combination of carefully engineered network-level and application-level features [27, 28, 65, 68]. We use similar insights to support discovery of disinformation websites based on infrastructure features. These features are particularly valuable because they are available before disinformation campaigns begin.

We construct features derived from a website’s domain, certificate, and hosting characteristics and then apply multilabel classification to categorize the website as disinformation, authentic news, or other (i.e., lacking news content). Our evaluation shows the feasibility of our approach on a large, labeled dataset of website snapshots. We also present results from a preliminary real-time deployment, in which we were able to discover previously unreported disinformation websites. We outline the challenges for automated disinformation detection based on our experience.

2 Definitions and Scope

Definitions of disinformation vary in academic literature and public discourse [37, 64]. Common components include intent to deceive about facts [18, 29, 34, 69], intent to harm [69], and intent to prompt distribution [40].

2.1 Websites as Granularity of Study

We study disinformation at the granularity of website domains, rather than individual articles, claims, advertisements, social media accounts, or social media actions (e.g., posts or shares). Websites are often used as distribution channels for disinformation on social media platforms [2, 3, 22, 24], and there are several benefits to studying disinformation at the website level:

- **Early Warning.** It is possible, in principle, to identify disinformation websites before they begin to publish or distribute content. For example, an automated detection
Figure 1: The lifecycle of a disinformation website, from domain registration to content distribution. We focus on features that are available as soon as the domain is registered, with progressively improving automated accuracy as the website deploys.

system might spot a new domain registration that looks like a local newspaper name, but has infrastructure overseas. A human moderator might then investigate and find there is no local newspaper with that name. Analysis of article content or social media activity, by contrast, can only occur much later in the disinformation lifecycle (Figure 1).

- **Longer-Term Value.** Disinformation articles and social media posts have an inherently limited lifespan, owing to the rapid news cycle [80]. Disinformation websites, by contrast, can last for years.

- **Platform Independence.** Identifying disinformation websites is feasible without access to a major online platform’s internal account or activity data.

- **Ecosystem Value.** A real-time feed of disinformation websites has value throughout the Internet ecosystem, similar to existing feeds of malware, phishing, and scam domains [20]. Website data is immediately actionable for a diverse range of Internet stakeholders. Further, because websites are often components of multimodal disinformation campaigns, detection at the domain level can provide an investigative thread to untangle the rest of a disinformation campaign, including associated social media accounts and activities.

There are drawbacks and limitations associated with focusing on domains. Some websites feature a mix of authentic and false news, complicating our class definitions (see Section 2.2). We also recognize that websites are just one source of disinformation, and that private communications and native social media content also play substantial roles in exposing users to disinformation and instilling false beliefs.

### 2.2 Class Definitions

**Disinformation** We define disinformation websites as websites that appear to be news outlets with content about politics and current events, but that operate in a manner significantly inconsistent with the norms, standards, and ethics of professional journalism. Satire websites fall within our definition of disinformation when the satire is not readily apparent to users. This is an intentional definitional decision, since satire websites can (and often do) mislead users [16, 35], and since disinformation websites are known to sometimes rely on implausible small-print disclaimers that they are satire [47, 48].

Our goal is to identify websites where users might benefit from additional context or other interventions. We decline to use the term “fake news”, even though it may be a more apt description of the category of website that we study, because of the term’s political connotations and because the utility of our features is generalizable.

We focus on websites related to current events and politics because a significant proportion of the U.S. population has encountered these types of disinformation websites at least once [2, 3, 24], and certain groups of users (such as those over the age of 65) encounter them at high rates [23].

**Authentic News** We define authentic news websites, including those with a partisan bias, as news websites that adhere to journalistic norms such as attributing authors, maintaining a corrections policy, and avoiding egregious sensationalism.

**Non-news** We define a third category of non-news websites, which are websites that primarily serve content other than news and do not represent (or claim to represent) news outlets.

### 3 Website Dataset

We used both current and historical data to construct our dataset. We identified three classes, rather than just two classes of fake and authentic news, both to facilitate feature engineering and because we found that cleaner class separation improved classification performance. We balanced the classes, including about 550 websites for each class.

We first constructed the disinformation class, then constructed the other two sets with equal sizes for balanced training and testing. Class sizes changed slightly over the course of dataset construction, so the final datasets contain 551 disinformation sites, 553 news sites, and 555 non-news sites. We recognize that the non-news website class would predominate in a real-time feed of domain, certificate, or social media events. Our rationale is that without balancing the dataset, the models we develop would minimize error by simply labeling every website as other.
3.1 Disinformation Websites

We began by combining multiple preexisting datasets of disinformation websites that had been manually labeled by experts or published by news outlets, research groups, and professional fact-checking organizations. Specifically, we integrated the corpora from CBS [10], FactCheck.org [17], Snopes [39], Wikipedia [71], PolitiFact [19], and BuzzFeed [61–63]. We also included websites that have been labeled as “disinformation” by OpenSources, a collaborative academic project that manually assigned credibility labels to news-like websites [79]. Finally, we integrated the list of disinformation websites compiled by Allcott et al. for their study on the diffusion of disinformation on Facebook between 2015 and 2018 [3], which they also compiled from lists by fact-checking organizations and academic sources.

We then manually filtered the list of websites, leaving only the websites that satisfied our definition of disinformation (Section 2.2). Our final dataset contains 758 disinformation websites. 575 (76%) of the websites are currently inactive: either unavailable, replaced with a parking page, or repurposed for other kinds of abuse (e.g., spam or malware distribution). This highlights the rapid turnover of disinformation websites in comparison to authentic news websites. Fortunately, we were able to reconstruct domain, certificate, and hosting features for 368 (64%) of these inactive websites through Internet Archive Wayback Machine snapshots, the DomainTools API, and the crt.sh Certificate Transparency log database [15, 33, 58]. This resulted in a set of 551 disinformation websites.

3.2 News Websites

We built a corpus of 553 authentic news websites, randomly sampling 275 from Amazon’s Alexa Web Information Service (AWIS) [4] and 278 from a directory of websites for local newspapers, TV stations, and magazines [75]. From AWIS, we sampled websites categorized as “news”, excluding the 100 most popular websites out of recognition that these websites likely have some distinct properties compared to the long tail of news websites (e.g., high-quality and customized infrastructure). From the local news dataset, we manually filtered to omit websites that did not prominently display news (e.g., TV station websites that served as channel guides).

3.3 Other Websites

We built a set of 555 other websites by sampling from Twitter’s Streaming API [66]. We filtered for tweets that contained a URL, extracted the domain name, and then used the Webshrinker classification service [70] to assign labels based on the Interactive Advertising Bureau’s standardized website categories [32]. We excluded websites that belonged to the “News” and “Politics” categories.

4 Feature Engineering

We engineered 33 features to distinguish disinformation, authentic news, and other websites (see Table 1). In this section, we describe exemplary features.

4.1 Domain Features

Eighteen features related to a website’s domain name, registration, or DNS configuration. Domain names were valuable for distinguishing authentic and disinformation websites from other websites; these classes generally used a domain name with a news-related keyword like “news”, “herald”, or “chronicle”. Domain registrars showed distinct usage patterns between authentic and disinformation websites: two low-cost and consumer-oriented registrars (Namecheap and Enom) were much more common among disinformation websites, while business-oriented registrars like Network Solutions, MarkMonitor, and CSC were more common for authentic news websites. We also found that authentic news websites were much more likely to have old domain registrations with far-off expiration dates, and that disinformation websites were much more likely to use new TLDs like .news. We observe that disinformation creators trying to evade detection based on domain features would need to plan in advance, establish business relationships, and bear other costs that may be deterring.

4.2 Certificate Features

Nine features related to a website’s TLS/SSL certificate. One key feature was the number of domains that a certificate covers (based on the Subject Alternative Name field). We found that news websites often had more domains in their certificates than disinformation websites, because parent news organizations used one certificate to cover their subsidiaries. We also found, though, that some disinformation websites had a large count of domains in their certificates attributable to low-cost hosting providers that deployed shared certificates.

4.3 Hosting Features

Six features pertained to a website’s hosting infrastructure. Using BGP routing tables, we identified the autonomous system hosting each website’s IP address and found that mass-market hosting providers like GoDaddy and Namecheap were much more common among disinformation websites, while premium, business-oriented hosting providers like Incapsula were more common among authentic news websites. We also geolocated IP addresses using the MaxMind GeoLite2 database and found that, contrary to our expectations, geolocation was not a valuable feature because most websites in all three classes were U.S.-based.

The presence of web trackers and low-quality advertisements may also be useful features, although we did not find that a feature based on websites’ Google Analytics IDs improved model accuracy. We leave the evaluation of other trackers to future work.
Table 1: Domain, certificate, and hosting features that our model uses to classify a website as authentic news, disinformation, or other. Features are ranked by Gini importance in our random forest model.

<table>
<thead>
<tr>
<th>Name</th>
<th>Category</th>
<th>Description</th>
<th>Rank</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>News Keyword(s) in Domain</td>
<td>Domain</td>
<td>The domain name contains one or more keywords that imply it serves news (e.g., “herald,” “tribune,” or “chronicle”).</td>
<td>1</td>
<td>Boolean</td>
</tr>
<tr>
<td>Domain Name Length</td>
<td>Domain</td>
<td>The number of characters in the domain name.</td>
<td>3</td>
<td>Numeric</td>
</tr>
<tr>
<td>“News” in Domain</td>
<td>Domain</td>
<td>The domain name contains the specific keyword “news.”</td>
<td>8</td>
<td>Boolean</td>
</tr>
<tr>
<td>WHOIS Privacy</td>
<td>Domain</td>
<td>The domain registrant is using a WHOIS proxy service or registrar privacy option.</td>
<td>9</td>
<td>Boolean</td>
</tr>
<tr>
<td>Registrar Name</td>
<td>Domain</td>
<td>The organization with whom the domain was registered.</td>
<td>11</td>
<td>Categorical</td>
</tr>
<tr>
<td>Nameserver SLD</td>
<td>Domain</td>
<td>The second-level domain of the nameserver.</td>
<td>14</td>
<td>Categorical</td>
</tr>
<tr>
<td>Nameserver AS</td>
<td>Domain</td>
<td>The autonomous system of the nameserver’s IP address.</td>
<td>16</td>
<td>Categorical</td>
</tr>
<tr>
<td>Registrant Organization</td>
<td>Domain</td>
<td>The organization of the registrant.</td>
<td>17</td>
<td>Categorical</td>
</tr>
<tr>
<td>Registrant Country</td>
<td>Domain</td>
<td>The country of the registrant.</td>
<td>19</td>
<td>Categorical</td>
</tr>
<tr>
<td>Time Since Domain Registration</td>
<td>Domain</td>
<td>The time elapsed since the domain was originally registered.</td>
<td>21</td>
<td>Numeric</td>
</tr>
<tr>
<td>Domain Lifespan</td>
<td>Domain</td>
<td>The time period between the domain’s initial registration and expiration dates.</td>
<td>22</td>
<td>Numeric</td>
</tr>
<tr>
<td>Time to Domain Expiration</td>
<td>Domain</td>
<td>The time until the domain’s registration expires.</td>
<td>23</td>
<td>Numeric</td>
</tr>
<tr>
<td>Time Since Domain Update</td>
<td>Domain</td>
<td>The time since the domain’s configuration was updated.</td>
<td>25</td>
<td>Numeric</td>
</tr>
<tr>
<td>Nameserver Country</td>
<td>Domain</td>
<td>The country where the nameserver is located, using IP geolocation.</td>
<td>27</td>
<td>Categorical</td>
</tr>
<tr>
<td>Novelty TLD</td>
<td>Domain</td>
<td>The TLD is novelty (e.g., .news, .xyz, or .club).</td>
<td>29</td>
<td>Boolean</td>
</tr>
<tr>
<td>Digit in Domain</td>
<td>Domain</td>
<td>The domain name contains numeric characters.</td>
<td>30</td>
<td>Boolean</td>
</tr>
<tr>
<td>Hyphen in Domain</td>
<td>Domain</td>
<td>The domain name contains a hyphen.</td>
<td>31</td>
<td>Boolean</td>
</tr>
<tr>
<td>Domain Resolves</td>
<td>Domain</td>
<td>The domain name resolves to an IP address.</td>
<td>32</td>
<td>Boolean</td>
</tr>
<tr>
<td>SAN Count</td>
<td>Certificate</td>
<td>The number of domains in the Subject Alternate Name extension field.</td>
<td>2</td>
<td>Numeric</td>
</tr>
<tr>
<td>SAN Contains Wildcard</td>
<td>Certificate</td>
<td>The Subject Alternate Name extension field contains a wildcard entry for a domain.</td>
<td>7</td>
<td>Boolean</td>
</tr>
<tr>
<td>Expired Certificate</td>
<td>Certificate</td>
<td>The certificate is expired.</td>
<td>10</td>
<td>Boolean</td>
</tr>
<tr>
<td>Certificate Available</td>
<td>Certificate</td>
<td>A certificate is configured at the domain (i.e., a certificate is provided during a TLS handshake on the HTTPS port).</td>
<td>12</td>
<td>Boolean</td>
</tr>
<tr>
<td>Self-signed Certificate</td>
<td>Certificate</td>
<td>The certificate is signed by the domain owner, not a CA.</td>
<td>13</td>
<td>Boolean</td>
</tr>
<tr>
<td>Domain-validated Certificate</td>
<td>Certificate</td>
<td>The domain owner has obtained a certificate from a CA through domain validation rather than organization validation or extended validation.</td>
<td>18</td>
<td>Boolean</td>
</tr>
<tr>
<td>Certificate Issuer Name</td>
<td>Certificate</td>
<td>The organization or individual who issued the certificate.</td>
<td>24</td>
<td>Categorical</td>
</tr>
<tr>
<td>Certificate Issuer Country</td>
<td>Certificate</td>
<td>The country where the certificate was issued.</td>
<td>26</td>
<td>Categorical</td>
</tr>
<tr>
<td>Certificate Lifetime</td>
<td>Certificate</td>
<td>The certificate’s period of validity.</td>
<td>28</td>
<td>Numeric</td>
</tr>
<tr>
<td>WordPress Plugins</td>
<td>Hosting</td>
<td>WordPress plugins used by the website.</td>
<td>4</td>
<td>Categorical</td>
</tr>
<tr>
<td>Website AS</td>
<td>Hosting</td>
<td>The autonomous system of the website’s IP address.</td>
<td>5</td>
<td>Categorical</td>
</tr>
<tr>
<td>WordPress CMS</td>
<td>Hosting</td>
<td>The website uses WordPress as its content management system.</td>
<td>6</td>
<td>Boolean</td>
</tr>
<tr>
<td>WordPress Theme</td>
<td>Hosting</td>
<td>The WordPress theme used by the website.</td>
<td>15</td>
<td>Categorical</td>
</tr>
<tr>
<td>Website Country</td>
<td>Hosting</td>
<td>The country where the website is located, using IP geolocation.</td>
<td>20</td>
<td>Categorical</td>
</tr>
<tr>
<td>Website Available</td>
<td>Hosting</td>
<td>A website is hosted at the domain (i.e., content is returned in response to an HTTP request for the base URL, following redirects).</td>
<td>33</td>
<td>Boolean</td>
</tr>
</tbody>
</table>

5 Model

We selected a multi-class random forest model because we expected that feature interactions would contribute to performance and that model interpretability would be important for evaluation and plausible deployment. We conducted a randomized hyperparameter search over a wide range of values, then selected values that achieved the best average accuracy based on 250 iterations with five-fold cross-validation.

Our model was able to distinguish the three classes and classify websites accurately; the mean ROC AUCs for authentic news websites, disinformation websites, and other websites were 0.98, 0.95, and 0.98 respectively, and the mean
Figure 2: Model performance on the curated dataset. Figures 2a and 2b show performance on all three classes using all feature types. Figures 2c and 2d show performance on the disinformation class using subsets of features.

Figure 2: Performance on 100 manually labeled websites from each class, sampled from real-time pilot data.

6 Pilot Real-Time Deployment

We conducted a pilot real-time deployment of our classifier to understand how well its performance generalizes. Our implementation used a commodity server to ingest new domains from DomainTools [15] (which tracks domain registrations), CertStream [9] (which tracks new TLS/SSL certificates), the Twitter Streaming API [66], and the Reddit API [55], then another commodity server to collect infrastructure data, generate features, and output a classification. We ran the system for 5 days, classifying 1,326,151 websites. To simulate the use of...
our classifier on a social media platform that has access to all features, we randomly sampled 300 websites that appeared on Twitter—100 from each detected class—and manually labeled to evaluate performance. We present a confusion matrix with classification performance in Figure 3.

Our model’s precision on the disinformation class (0.05) was sufficient for plausible deployment. We were able to rapidly discard false positives and, just in our small-scale pilot deployment, discovered two disinformation websites that had not been previously reported in any public venue.

We cautiously note that model performance radically degraded in comparison to our prior evaluation, similar to results in concurrent work [12]. We attribute this difference to three potential causes. There is a massive class imbalance inherent in real-world classification—the overwhelming majority of websites do not relate to news, and most news links shared on social media are for authentic news websites rather than disinformation websites. As a result, even with good performance, false positives may dominate true positives. Also, as training data gets older, the features for current websites may change. This may create a gap between the features observed in the training data and current disinformation websites.

It is also possible that our model is picking up on artefacts in our training dataset to classify websites. For example, ∼34% of the disinformation websites in our training data are active, whereas all of the news websites are active. If there are differences in the features of inactive websites that we reconstructed features for and active websites, then our classifier may be using the wrong signal to distinguish disinformation websites from news websites. We leave an evaluation of our classifier that is only trained on active websites to future work.

7 Evasion Resistance

Disinformation website operators will be motivated to evade detection. In other areas of online abuse, such as spam, phishing, and malware, online platforms are constantly developing new defensive measures to keep up with advances in adversary capabilities. We expect that disinformation will follow a similar cat-and-mouse pattern of defense and evasion.

Our model uses features that provide a degree of asymmetric advantage in identifying disinformation websites, since a website that seeks to evade detection must make changes to its infrastructure. Some features will be relatively easy to evade; for example, a website can easily change a WordPress theme or renew an expired TLS certificate. Fortunately, many of the most important features that our model relies on are difficult or costly to evade.

As an example, consider one of the most predictive features: the lifespan of a website’s domain. Evading that feature requires either significant advance planning or purchasing an established domain. Evading certain other features incurs monetary costs, like purchasing a certificate from a reputable issuer, registering a domain for a longer time, switching to a more expensive non-novelty TLD, or migrating to a more trustworthy hosting provider. Evading other features incurs technical costs: obtaining and installing a correctly configured, reputedly issued TLS certificate, for instance, imposes some operational cost. Finally, evading many of our model’s features might reduce the effectiveness of the disinformation campaign. For example, a top ranked feature is whether a domain contains news keywords. Removing those keywords from the domain name could diminish the credibility of the website and lead to less exposure on social media.

8 Related Work

Our approach is inspired by the information security literature on detecting malicious websites. Prior work has demonstrated the value of infrastructure features for identifying spammers [28, 52], botnets [21, 53], and scams [26, 27, 36]. Efforts at classifying disinformation have predominantly used natural language features (e.g., [1, 11, 13, 30, 38, 42, 44, 46, 49–51, 54, 56, 57, 60, 72, 77]). Several projects have relied on social graph features (e.g., [25, 41, 73, 74, 78]).

In the prior work closest to our own, Baly et al. predicted news website factuality by examining the domain name, article text, associated Wikipedia and Twitter pages, and web traffic statistics for a combined accuracy of about 0.5 [6]. Concurrent to our work, Chen and Freire developed a system for discovering political disinformation websites on Twitter that uses article text and markup [12]. The authors report a mean ROC AUC of 0.97 on historical data and significantly lower performance in a trial real-time deployment.

In comparison to the disinformation detection literature, we contribute a new set of infrastructure features that do not rely on content or distribution. We demonstrate website classification with state-of-the-art performance and characterize how performance degrades in a real-world deployment setting.

9 Conclusion

Our work demonstrates a promising new set of features for detecting disinformation websites and a real-time classification system could feasibly be deployed by online platforms. Future work on disinformation detection should examine how infrastructure features interact with natural language and social sharing features. Our work also highlights the significant and unexplored practical challenges of real-world deployment for disinformation detection systems. As a research community, we have an opportunity to support the free and open speech environment online by enabling responses to disinformation. But we have to move beyond detection methods that are only strong on paper and begin addressing the hard problems associated with real-world deployment.

Acknowledgments

We thank our paper shepherd Jed Crandall and our anonymous reviewers for their feedback. This work was funded in part by National Science Foundation Award TWC-1953513.
References


Appendix

A.1 Domain Proxy Keywords

We first computed the most popular values of the “registrant organization” field from WHOIS records in our training data. Then, we manually extracted keywords from known domain proxy services that were highly ranked. We used the resulting list of keywords as a heuristic to determine if new domains use WHOIS privacy services:

- domain
- protect
- whois
- guard
- proxy
- privacy
- redacted

A.2 News-Indicating Keywords

We used the following 163 keywords with the DomainTools Brand Monitor API to retrieve newly registered domains that may be news websites. We derived the keywords by manually examining our training datasets of authentic and fake news websites for common terms related to news, media, information, or publishing.

```
24 365 abc action
activist advance alert alliance
alternative america associate blast
blog box breaking brief
bulletin business buzz byte
caller cbs channel christian
chronicle citizen city club
cnn conservative corner county
courier current daily democracy
dig dispatch division edition
diary election empire epoch
evening examiner express extra
fact feed file finesser
fresh gazette global guardian
hangout headline herald hq
hub idea independent index
info inquire insider interesting
international item journal leak
learn ledger liberal liberty
live local mag maga
mail media metro movement
nation nbc network now
observer page paper patriot
pioneer pipe plug politic
post press progress proud
publish radio react read
record region religion report
republik review rumor scoop
sentinel share sharing show
spirit spot spotlight standard
star state statesman stories
story studio sun surge
syndicate telegram telegraph television
times today top trend
tribune truth tv twenty-four
twentyfour uncut underground union
update us usa view
viral vision washington watch
weekly wire your zine
zone
```