Social Network Analysis and Mining (2024) 14:153
https://doi.org/10.1007/s13278-024-01330-w

ORIGINAL ARTICLE

Can computer network attributes be useful for identifying low-credibility websites? A case study in Brazil

João M. M. Couto[1] · Julio C. S. Reis[2] · Fabrício Benevenuto[1]

Received: 31 May 2024 / Revised: 31 July 2024 / Accepted: 1 August 2024

Abstract
Misinformation has become a global issue with impacts on various spheres of society, particularly in developing countries like Brazil. In most misinformation ecosystems, a recurring challenge is the spread of fake news through websites that closely replicate the look and function of reputable news outlets. This facilitates their dissemination, which might also involve automation, political bias, targeted ads and even the exploitation of social network algorithms in an attempt to reach niche audiences. Due to this high complexity and the rapidly evolving nature of the problem, we are just beginning to identify patterns in the various misinformation ecosystems on the Web. In this work, we extend a previous study, offering important steps towards a deeper understanding of computer networking patterns observed on Brazilian misinformation websites. These patterns emerge from various sources, including DNS records, domain registrations, TLS certificates and hosting infrastructure. Our results reveal a novel avenue through which low-credibility news websites can be distinguished from websites of credible nature.

Keywords Misinformation · Fake news · Credibility · Brazilian websites · Computer network attributes

1 Introduction

Misinformation campaigns on the Internet have affected several countries in recent years, endangering the integrity of public discourse, electoral processes, and democratic governance (Allcott and Gentzkow 2017; Spohr 2017; Ferrara 2017; Lazer et al. 2018). In the Coronavirus pandemic, for instance, the issue has reached new levels, including the diminishment of public health concerns, promotion of medication without proven efficacy, or dismissal of sanitary measures (van Der Linden et al. 2020; Depoux et al. 2020). Authorities have recognized the problem worldwide: in 2021, the Nobel Peace Prize was awarded to two internationally recognized journalists for their efforts in the fight against misinformation and the defense of freedom of speech.[1]

Particularly in Brazil, the 2018 presidential elections were a stage for an expansive distortion of the truth promoted by misinformation campaigns launched by entities with well-defined agendas, notably in messaging apps such as WhatsApp (Resende et al. 2019; Melo et al. 2019). In this context, misinformation has gained an unprecedented magnitude in the country, greatly empowered by digital platforms (e.g., social networks, messaging apps, etc). The observed effectiveness of those campaigns now brings about widespread concern that this phenomenon will recount itself in the subsequent presidential elections and other related events.

Organized efforts to produce and disseminate misinformation are increasingly complex, exploiting a whole array of digital platforms as well as different techniques to artificially maximize the reach of this type of content (Silva et al. 2020; Ferrara et al. 2016; Ribeiro et al. 2020). In this context, a vital component of most misinformation campaign consists of deploying websites that publish news in a similar fashion to news articles published by credible sources, but containing fake stories, often associated with sensitive subjects, such as politics.

In this work, we tackle the problem of measuring and understanding the characteristics of these kinds of

✉ João M. M. Couto
joaocouto@dcc.ufmg.br

[1] Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil

[2] Universidade Federal de Viçosa (UFV), Viçosa, Brazil

[1] https://www.nytimes.com/live/2021/10/08/world/nobel-prize

Published online: 07 August 2024

websites. Particularly, we focus on characterizing networking attributes, which, to the best of our knowledge, are not explored in previous studies. We build a set of low-credibility Brazilian websites by finding at least one news piece whose veracity can be directly contested through an article published by a recognized fact-checking agency. For comparison, we also build a list of high-credibility news outlets in Brazil. We then gather publicly available data for all these websites, including information from DNS records, IP address and domain name registrations, TLS certificates, and hosting infrastructure. Finally, we explore this information to extract attributes and characterize these websites dedicated to disseminating misinformation through digital platforms in Brazil.

Our analysis unveils valuable patterns. We show that active low-credibility websites are often registered just recently and do not present TLS certificates and domain name expiration dates as long-lived as those of high-credibility websites. We also find that low-credibility websites are often registered abroad, bypassing Brazil's registration system (which requires personal identification), making them more resilient against takedowns. We hope our study can be useful to help authorities and interested institutions (e.g. fact-checking agencies) to identify Brazilian low-credibility websites.

In particular, this work adds new contributions to our previous effort (Couto et al. 2022). First, we contextualize our work by presenting existing solutions and their limitations, some of which we attempt to overcome with this work. Second, we provide a more in-depth characterization of our proposed attributes highlighting further evidence of their individual contributions to the discrimination of high and low-credibility websites. Last, we also demonstrate the joint discriminative power of the attributes explored in this work, first through an intuitive T-SNE visualization and later through a machine learning algorithm and its associated performance metrics, which emphasize their potential to distinguish low-credibility websites from others.

The rest of the paper is organized as follows. In Sect. 2 we discuss related works. Section 3 describes the creation process of the low- and high-credibility websites datasets. An overview of the proposed and computed networking attributes is presented in Sect. 4. Then, in Sects. 5, 6, and 7 we detail the results, presenting the main differences between the high- and low-credibility websites in the Brazilian context and their potential to distinguish them. Finally, Sect. 8 concludes our work and offers some directions for future ones.

2 Related work

The research interest from different areas in the misinformation phenomena on digital platforms is growing. There has been a great effort by the computing community towards *(i)* understanding (Shu et al. 2018; Lazer et al. 2018; Vosoughi et al. 2018; Resende et al. 2019) and *(ii)* proposing solutions that are effective against the problem (Reis et al. 2019; Rath et al. 2022; Bazmi et al. 2023) considering different platforms such as Twitter (Helmstetter and Paulheim 2018) and YouTube (Hussain et al. 2018). With a focus on *(i)* providing a better understanding of news propagation in the misinformation ecosystem, Fletcher et. al (2018) analyzed the spread and reach of particular misinformation instances in France. This effort revealed that the most popular misinformation websites had reached 1.5 million French citizens (Fletcher et al. 2018). Lastly, more closely related to our work, Bozarth and Budak (Bozarth and Budak 2020) investigate the role of ad servers for low- and high-credibility websites in the USA and results revealed that fake news domains more extensively employ ad servers and that those ads are much more likely to be risky in nature.

Furthermore, recently there have been various studies to develop mechanisms capable of differentiating misinformation from factual information. Studies in this direction *(ii)* commonly make use of machine learning strategies and statistical modeling to create classifiers capable of identifying misinformation through textual patterns extracted from specific instances of misinformation content (Pérez-Rosas et al. 2017; Volkova et al. 2017) or contextual information regarding the propagation of content on social media (Reis et al. 2019). On the other hand, there are some efforts to perform automatically misinformation detection comprising works that, for monitoring online misinformation, aim at exploring tools or online systems (Giełczyk et al. 2019). These systems were proposed and used as countermeasures to the misinformation phenomena on different digital platforms. Examples include "Fake tweet buster" (Saez-Trumper 2014), a Web tool to identify users promoting fake news on Twitter, Hoaxy (Shao et al. 2016), a Web platform for the tracking of social news shared containing misinformation, WhatsApp Monitor (Melo et al. 2019), Facebook Ads Monitor (Silva et al. 2020), a system for auditing the political use of ads in Facebook, and Telegram Monitor (Júnior et al. 2022), which displays the most popular content shared on public groups in WhatsApp and Telegram, respectively, helping fact-checking agencies to identify relevant content to check.

Despite the undeniable importance of these efforts, they are limited. Most of them explore ways to combat misinformation by focusing on specific misinformation instances (e.g., news stories later labeled by a fact-checking agency),

Table 1 Overview of the datasets

Low-credibility	71 websites
High-credibility	98 websites

or by using attributes that are frequently only available for analysis after news instances have already been propagated on a given digital platform – such as Twitter and Facebook (e.g., share count of a news item or sharing patterns). Thus, in this work, we aim at complementing previous efforts that enable a better understanding of the problem. We offer a significant differential by providing a measurement of network attributes related to low-credibility news websites, which, to the best of our knowledge, are not explored in previous efforts. Note that computing network attributes are not restricted to specific misinformation instances and are available as early as the creation of a new website. As such, we believe that characterizing the patterns associated with these attributes can be useful for constructing effective mechanisms for the early identification of (low credibility) websites that generate and host misinformation content.

3 Dataset construction

Ideally, we would like to have at our disposal a curated list of low- and high-credible Brazilian websites. However, in Brazil, a list of low-credible websites is quite hard to be obtained. Fact-checking agencies avoid explicitly pointing to the sources of debunked claims as oftentimes it leads to legal backlash and expensive litigation from actors behind misinformation campaigns.[2] Thus, unfortunately, such a list is not available, compelling us to construct one from scratch. In this section, we briefly describe our strategy to build a dataset of low- and high-credibility websites, which are presented in Table 1.

3.1 Low-credibility websites

To identify low-credibility websites, we propose a strategy based on the hypothesis that a user or account (maintained by a person or a bot) posting a piece of misinformation on a digital platform (e.g., Twitter) is likely to post additional ones.

First, manually identify an initial "*seed*" consisting of a news article containing misinformation. Critically, this article must have been assessed by an internationally recognized fact-checking agency and confirmed as misinformation or fake news. In our work, we only consider assessments provided by agencies that are signatories of the International Fact-Checking Network (IFCN),[3] which sets guidelines and best practices for fact-checking. With this seed, we use a Twitter data collector, implemented with Python and Selenium, to navigate the results of searches for the seed and identify all users who shared it. The collector downloads each user's timeline, including tweets and replies, from which those not containing links are excluded. Then, we extract all website domain names from these links. The top 20 most tweeted links from these domains are then considered candidate misinformation instances. This constant was determined through the analysis of the average popularity ranking obtained by labeled misinformation instances (Araujo et al. 2024).

For each candidate, we identify 3–5 keywords from their titles (Araujo et al. 2024) by first removing stop words and then identifying noun and verbs that carry the most meaning: proper nouns (names, places, organizations) are particularly important, as well as the subset of verbs associated with them. Finally, we rank all fact-check articles from a fact-check dataset by the presence of those keywords in their titles and bodies. This fact-check dataset contains every fact-check published by Brazilian IFCN agencies and is available at: https://doi.org/10.5281/zenodo.5191798 (Marques et al. 2022). Finally, for each candidate we manually verify whether the top-ranking fact-checks directly assess the veracity of its claims: if they do, the candidate is labeled a confirmed instance of misinformation. The misinformation instance (i.e., articles with a corresponding fact-check article directly attesting its falsehood) can then be used as input in the same fashion as the original misinformation seed. As such, the Twitter data collector iteratively searches for additional users tweeting misinformation and additional candidate phony news articles.

In this work, we executed 7 iterations of this process. In total, we identify 71 websites that published at least one news article containing misinformation, which we label as low-credibility websites. The initial misinformation seed that we used to bootstrap our search asserts the coronavirus as a biological weapon,[4] which was later contested by a fact-check from the agency *Estadão Verifica*.[5]

[2] https://revistaoeste.com/brasil/agencia-de-checagem-aos-fatos-e-condenada-por-publicar-fake-news/ (in Portuguese)

[3] https://www.ifcncodeofprinciples.poynter.org/

[4] https://www.estudosnacionais.com/32878/novo-coronavirus-pode-ter-sido-fabricado-como-arma-biologica-claims-researcher/ (in Portuguese)

[5] https://politica.estadao.com.br/blogs/estadao-verifica/coronavirus-site-distorce-entrevista (in Portuguese)

3.2 High-credibility websites

We consider high-credibility websites to be those accredited by Brazil's National Association of Newspapers (ANJ), a nonprofit organization recognized internationally for its responsibility in misinformation prevention. ANJ's statute[6] mandates that all members comply with its code of ethics, which emphasizes verifying and publishing the truth of facts of public interest. This dataset serves as a baseline to verify if there are significant differences, based on presented attributes, distinguishing low-credibility websites from others.

A total of 98 websites from this list were included as high-credibility sites in this work. We applied the same process of matching news titles with fact-checks as done for the low-credibility database. Starting with an initial news article, we ensured none of the high-credibility website articles matched any fact-check titles. Similarly, we verified that none of the low-credibility websites were accredited by ANJ.

4 Attribute categories

The deployment and maintenance of a website on the Internet require the acquisition of resources as well as effort to configure the hosting infrastructure. The specifics of the resources employed by a website may vary, for example, based on goals, available funding, target availability, popularity, user base expectations, and technical staff expertise.

In this work, we collect attributes associated with resources used by low- and high-credibility websites in an attempt to identify significant differences between them. Thus, we implement tools to collect 31 publicly-available website attributes clustered in the three classes discussed below. An important property is that all attributes are publicly available as soon as a website is created, thus they present the potential for monitoring systems to quickly flag low-credibility websites.

In addition, all the attributes can easily be obtained through publicly available query protocols (e.g., WHOIS[7] (Daigle 2004)) or through low-cost commercial tools (e.g., IPStack,[8] used here to geolocate the different websites).

Table 2 offers an overview of the computed attributes that can be grouped into three sets: Domain, Certificate, and Geolocation, which are described next.

4.1 Domain

The resolution of domain names via the DNS system (*Domain Name Service*) is essential for most public websites. As such, a domain's registration and the configuration of its authoritative DNS servers provide essential information about its nature. Moreover, they enable the identification of patterns that serve as indications of the infrastructural robustness of a given website and possibly the registrant's intentions when a new domain is created.

An intuitive example of this concept is the total duration a website's registrant decides to purchase a domain name for. At the purchase of a domain associated with a new misinformation campaign, it is reasonable to expect that registrants are aware of the possibility that at any given point, a court order might take them down. Therefore, this type of website has a greater tendency to observe shorter-term registration renewals as a way to avoid significant financial losses. As detailed later in the results section, the observed days-until-expiration values of low-credibility websites are much lower than that of high-credibility sites.

Similarly, other attributes associated with the domain, such as the use of privacy options for the WHOIS protocol or the nature of the registrar utilized to purchase the domain, are also indications that might be associated with the registrant's intentions and a domain's purpose. Here, we implemented 14 attributes of this set (i.e., Domain). In sum, these attributes are associated with the registration and configuration of the domain name, including DNS data.

4.2 Certificate

Support for encrypted access to public websites is an increasingly adopted practice on the Internet, mainly through the use of the *HyperText Transfer Protocol Secure* (HTTPS) protocol where websites would previously employ *HyperText Transfer Protocol* (HTTP). Offering HTTPS on a website involves issuing TLS certificates necessary to authenticate the identity of servers responding to requests directed to a domain.

Similarly to DNS, we can extract attributes that might correlate with the nature of the different websites analyzed. In particular, most TLS certificate generation and maintenance services incur recurring costs, often proportional to the support, robustness, or optional certificate properties.

This phenomenon manifests itself in different ways. As an illustration, we have highlighted the total validity period of the certificate: free TLS certification services generally issue certificates valid for less than or equal to 90 days, while paid services frequently issue certificates that are valid for a year or more. In this scenario, for reasons similar to those presented in the context of a domain purchase in the previous subsection, that is, to avoid more significant financial

[6] https://bit.ly/3NFFb9O (in Portuguese)

[7] Protocol for querying registration information associated with entities on the Internet: https://who.is.

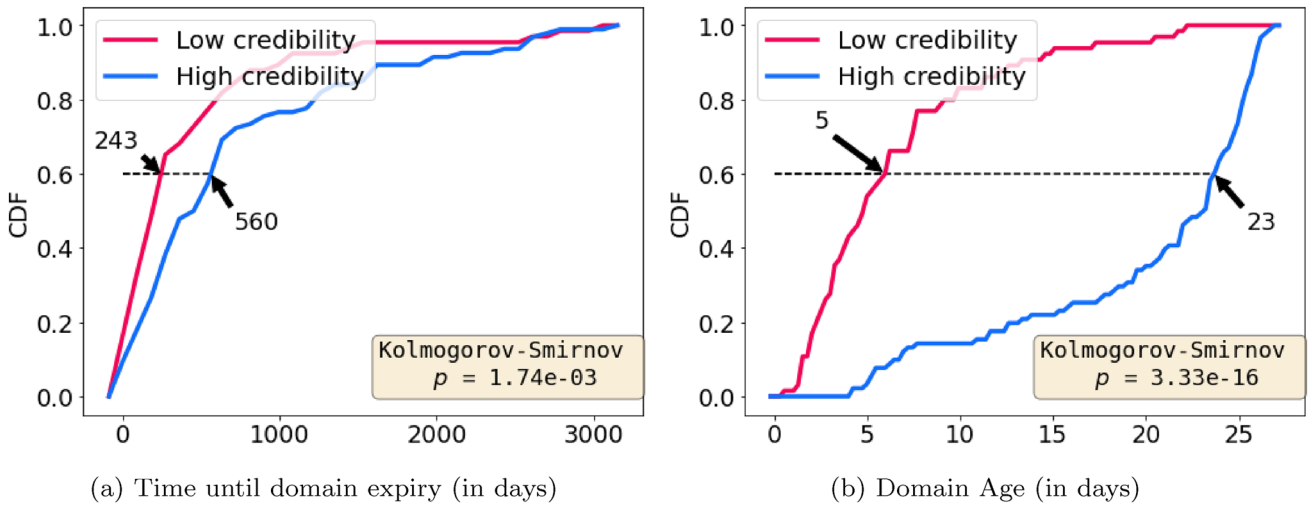
[8] https://ipstack.com/

Table 2 Extracted attributes aggregated by category (i.e., Domain, Certificate and Geolocation attributes) and data type (i.e., Boolean, Num. = Numerical, Cat. = Categorical)

Identifier	Data type	Description
Domain attributes (D)		
Subdomain-hifen	Boolean	Subdomain contains a hyphen
Subdomain-digit	Boolean	Subdomain contains a digit
Tld-br-or-com	Boolean	TLD (Top Level Domain) is either.br or.com (most common TLDs in the dataset)
News-keywords	Boolean	URL contains a journalistic keyword (e.g., gazette, tribune)
Whois-privacy	Boolean	Registrant enabled WHOIS privacy options
Resolution-hops	Numerical	Number of hops necessary to resolve subdomian into a IP address
caa-txt-count	Numerical	Number of CAA or TXT entires in the domain's DNS
len-subdomain	Numerical	Number of characters in the subdomain
Domain-age	Numerical	Time, in days, since the initial registration of the domain
Domain-expiry	Numerical	Time, in days, until the domain expiry date
Domain-update	Numerical	Time, in days, since the domain's DNS was last modified
as-n	Categorical	Autonomous System Number associated with the subdomain's IP
Registrar	Categorical	Registrar utilized for registering the domain (e.g., GoDaddy)
Registrar-url	Categorical	Registrar URL (e.g., NameCheap.com)
Certificate attributes (C)		
Allows-http	Boolean	Domain is acessible via HTTP requests
Redirects-http	Boolean	Server redirects requests HTTP
ca-is-letsencrypt	Boolean	Certificate Issuer is the popular "Let's Encrypt" free certification service
cert-expired	Boolean	TLS certificate has expired
Public-key-bits	Numerical	Number of bits present in the domain's public key used for TLS handshake
cert-age	Numerical	Time since the TLS certificate was issued (days)
cert-expiry	Numerical	Time until the expiry date of the TLS certificate (days)
cert-lifespan	Numerical	Total TLS certificate lifespan (issuing until expiry)
ca-entity	Categorical	TLS certificate issuing entity
ca-nacionality	Categorical	Country code associated with the TLS certificate issuing entity
Geolocation attributes (G)		
ip-in-brazil	Boolean	Geolocation of IP address resulted in coordinates within Brazil
ip-in-usa	Boolean	Geolocation of IP address resulted in coordinates within the U.S
as-ip-equal-cc	Boolean	IP address coordinates within the country associated with ASN
ip-cc	Categorical	Country code associated with the IP address coordinates
as-cc	Categorical	Country code associated with the ASN
lat-ip	Numerical	Geolocated IP address latitude
Long-ip	Numerical	Geolocated IP address longitude

losses, it is reasonable to expect that low credibility sites will, more likely, employ low (or no) cost TLS certification services. We will see that in our database, only a tiny portion of certificates from low-credibility sites will expire in more than one year after the data was collected, meanwhile that is the case for nearly a quarter of the high-credibility sites' certificates.

Other certificate properties may be correlated with the nature of domains. A TLS handshake initiates a TLS connection between a client (e.g., Web browser) and the server (e.g., website) during HTTPS. This process sees both ends of the connection verifying each other's identities, negotiating which encryption protocols will be employed before finally exchanging session keys. The number of bits present in the public key used during the *handshake* between web browsers and the server of a given news portal is another indicator of the robustness of the certification service employed by the website. In total, we implemented 10 attributes of this set (i.e., Certificate). In short, these attributes are related mainly with the certificate (or lack thereof) used by a website.



(a) Time until domain expiry (in days) (b) Domain Age (in days)

Fig. 1 Comparison of the two most dissimilar numerical attributes between high- and low-credibility websites

4.3 Geolocation

In Brazil, recurrent government initiatives aim to eliminate misinformation vectors.[9] Initiatives of this kind gain easy access to a range of legal tools that can be used to carry out quick shutdowns of websites or pages on social networks. Therefore, measures to make these processes more difficult become objects of interest for entities seeking to launch misinformation campaigns. In this context, the use of domain registration and hosting services offered by foreign entities represent one of the most popular measures. By effectively operating under the jurisdiction of another country, low-credibility websites significantly increase the legal complexity of proceedings aimed at their dismissal. In this context, the geolocation of a website becomes an object of interest that might contain useful information to help differentiate low-credibility news sources from their high-credibility counterparts. In this context, from the Geolocalization set, we implemented several attributes derived from the IP address hosting a website and its associated Autonomous System.

5 Websites characterization

In this section, we present the results for the three sets of attributes presented in Sect. 4. Our aim is to characterize both credibility groups and investigate the discriminative capacity of the extracted attributes. To this end, we demonstrate that the attributes calculated on the set of low credibility websites follow fundamentally different distributions from websites in the high credibility set.

For each numerical attribute (identified in Table 2) we report the mean, 60th percentile, and the p-value of the Kolmogorov-Smirnov (KS) test (Massey 1951). The KS test compute the probability that two samples come from the same underlying distribution, which we use to compare the distributions of attributes for the high- and low-credibility websites. If the p-value associated with a test is less than a significance level (we will use 0.05, or 5%) we expect that the attribute helps differentiate between high- and low-credibility sites, *i.e.,* they can serve as input attributes or features to a classifier model.

For categorical and Boolean attributes, we will present the incidence, in percentage points, of each category (false or true in the case of Boolean attributes) observed in the two populations. Similar to numerical attributes, categorical attributes with very different incidences can be useful to differentiate between high and low-credibility sites.

In the following subsections, we present the results of the two numerical attributes with the most dissimilar distributions for each attribute category (Domain, Certificate, and Geolocation, the incidence of categorical and Boolean attribute classes, as well as observations regarding remaining attributes.

5.1 Domain attributes

A domain name's registration attribute values (e.g., registrant, age, expiration date) can be obtained by querying WHOIS servers. We obtain configuration attribute values regarding the authoritative server associated with a domain name (e.g., number of CAA/TXT entries) by performing DNS queries through the `dig` command. In both cases,

[9] https://tinyurl.com/mr3uws8d (in Portuguese)

Table 3 Class distribution for categorical attributes by credibility group (LC = Low Credibility, HC = High credibility). Others includes CN, SC, AR, NL, RU country codes

		Associated country code							
		US	BR	GB	BE	CA	DK	DE	Others
IP	HC (%)	53.1	45.8	0.0	0.0	1.0	0.0	0.0	0.0
	LC (%)	80.3	13.6	0.0	0.0	0.0	1.5	0.0	4.5
ASN	HC (%)	60.4	37.5	0.0	0.0	1.0	0.0	1.0	0.0
	LC (%)	83.3	10.6	0.0	0.0	1.5	1.5	1.5	1.5
TLS	HC (%)	86.0	2.3	4.7	5.8	0.0	0.0	0.0	1.2
	LC (%)	91.7	0.0	6.7	1.7	0.0	0.0	0.0	0.0

query results are filtered to extract fields of interest through Python scripts that take into account the different response formats for these queries, which may vary depending on the TLD under which a certain domain was registered.

Time until domain expiration The expiration date of a domain is a function of the total registration period paid in advance. In this work, we investigated whether low-credibility sites hire registration services for a shorter duration when compared to high-credibility sites.

Figure 1a corroborates with that possibility: 60% of low credibility websites expire within 243 days while the same value for high credibility websites credibility is 560 days. The p-value calculated in the Kolmogorov-Smirnov test shows that the distributions of expiration times are different. As such, the time until the expiration date of a domain can be used as a contributing factor in the characterization of low-credibility websites.

Domain age The age of a website can be seen as an indicator of the continuity and consistency of efforts dedicated to maintaining and creating content for said website. Here, we intend to find out whether low-credibility websites tend to have been established more recently than high-credibility websites considering they might have a shorter service life as they frequently seek to meet the ephemeral agendas (*e.g.,* elections) or are more frequently taken down by court orders.

Figure 1b strongly supports this argument: 76% of low credibility sites are less than 5 years old. Websites within the high credibility set have a much longer lifespan: only 6.7% of them are less than 5 years old. The Kolmogorov-Smirnov test resulted in a near-zero p-value, indicating that this attribute behaves differently between the two sets of websites and therefore can be useful to distinguish them. One advantage of this property for discriminating between high- and low-credibility websites is that it cannot be easily faked as domain creation dates are maintained by registrars; when necessary, the registration date and hosted content can be matched by querying historical Web archives.

Categorical and boolean domain attributes In Table 3 we can observe the large discrepancy of the categorical attribute `ip-cc` (country of IP geolocation): only 13.6% of low credibility sites are hosted in Brazil, while the incidence among high credibility sites is 45.8%. This result indicates that low-credibility websites are operated abroad significantly more often than high-credibility websites.

Table 4 Occurrence of true label per credibility group and boolean attribute

		Boolean attributes true label presence	
		Low credibility(%)	High credibility (%)
D	Subdomain-hifen	9.9	2.0
	Subdomain-digit	15.5	6.1
	tld-br-or-com	91.5	98.0
	News-keywords	22.5	49.0
	Whois-privacy	12.7	8.2
C	Allows-http	15.5	12.2
	Redirects-http	80.3	78.6
	ca-is-letsencrypt	66.2	64.3
	cert-expired	15.5	14.3
G	ip-in-brazil	19.7	46.9
	ip-in-us	81.7	54.1
	as-ip-equal-cc	91.5	91.8

Among the Boolean attributes presented in Table 4, three stand out: the presence of a hyphen in the subdomain (9.9% vs 2.0%), the presence of digits in the domain (15.5 % vs. 6.1%), and the presence of journalistic keywords in the subdomain (22.5% vs. 49.0%). Furthermore, the resulting incidence difference in the `tld-br-or-com` attribute indicates that low-credibility websites may be more likely to use unusual TLDs (not.br or.com). While 98% of high-credibility sites are registered under the `.com` and `.br` TLDs, the figure for low credibility websites is a lower 91.5%.

Table 5 Certificate lifespan distribution

		Lifespan (in days)		
		90	365	>365
Certificates	High credibility (%)	41.7	34.5	23.9
	Low credibility (%)	52.8	44.9	2.4

5.2 Certificate attributes

To evaluate our intuition that the credibility of websites can be captured by the TLS certificate attributes proposed in Sect. 4.2, we extract certificate attributes via commands offered by the OpenSSL[10] library and also the `curl` command-line tool.[11] In particular, OpenSSL was used for extracting attributes associated with the certificates themselves, such as the issuance and expiration dates, and `curl` was used to determine whether a website accepts HTTP requests and if these types of connections are automatically redirected to HTTPS.

Lifespan of certificates The duration of a TLS certificate associated with a website is a function of the type of TLS certification service contracted by each website. In this context, we investigated whether certificates issued to highly credible sites have a longer validity period than those issued to low-credible websites. We observed that all certificates had a duration that is a multiple of 3 months, resulting in a numerical attribute that is only ever one of a few possible values, thus, its presentation is better suited in the form of an incidence table.

Table 5 presents the distribution of the different certificate lifetimes in the database. Notice that while 23.9% of certificates from highly credible sites last for more than one year, this is the case for only 2.4% of low-credibility websites. This result indicates that certificates of longer duration, in particular with a duration of more than one year, is indicative of high-credibility websites. In this context, it is important to note that *Let's Encrypt*, one of the most popular free services for issuing TLS certificates, only issues certificates with a duration of 3 months.[12]

Categorical and boolean certificate attributes The C section of Table 4 provides the incidence of each Boolean certificate attribute calculated on the two sets of websites. Here, it is notable that the attributes provide limited, if any, discriminating power between the credibility groups as the incidence is quantitatively similar.

[10] https://www.openssl.org/
[11] https://curl.se/
[12] https://letsencrypt.org/2015/11/09/why-90-days.html

5.3 Geolocation attributes

As aforementioned, we use the IPStack geolocation API[13] to obtain the geographic coordinates associated with the IP address each website's domain name resolved to. Figure 3 shows the location of websites, where color encodes credibility and point sizes capture the number of sites in the region. The map intuitively suggests the correlation between the coordinates obtained and the presence of misinformation content. It is easy to observe that the coordinates of the low credibility sites are more geographically distributed and more intensely present abroad, as we will quantify next.

Geographical concentration Many of the main hosting services for websites are located in major technological hubs. Figure 3 suggests that low-credibility websites have a greater tendency to use alternative services outside these centers. To verify this hypothesis, we translated the coordinates of each website into the corresponding municipality: we observed that the set of 98 high-credibility websites are hosted in services across 20 different cities, whereas the 71 low-credibility sites are spread in 28 different cities, suggesting that low-credibility websites are less concentrated in large hubs.

Pairs of distances To explore the intuition proposed in Sect. 4.3, we compute the geographic distance between pairs of websites in each credibility group. This metric allows us to compare if high or low-credibility websites are more or less geographically concentrated. To calculate these distances, we use the Vincenty distance implemented in the Geopy library.[14] For each website, we calculate the distance between its coordinates and the coordinates of all other websites in its credibility group.

Figure 2a presents the distribution of the distances in the two groups, indicating the 60^{th} percentile and the p-value obtained in the Kolmogorov-Smirnov test. Here, it is important to note that low-credibility websites are more heavily concentrated in the US than high-credibility sites are concentrated in Brazil. Thus, low-credibility websites generate pairs of smaller distances despite being present in a greater number of countries around the globe. In particular, the mode of approximately 10,000 km distance between high-credibility websites results from several of them being hosted either in São Paulo, Brazil, or the Bay Area; while a large number of distances below 4000 km among low-credibility websites captures that most of them are located within the US (see Table 3).

The p-value obtained in the Kolmogorov-Smirnov test between the two sets of distance pairs was much lower than 5%, demonstrating the divide between the distributions observed for this attribute in the credibility groups.

[13] https://ipstack.com
[14] https://geopy.readthedocs.io/en/stable/

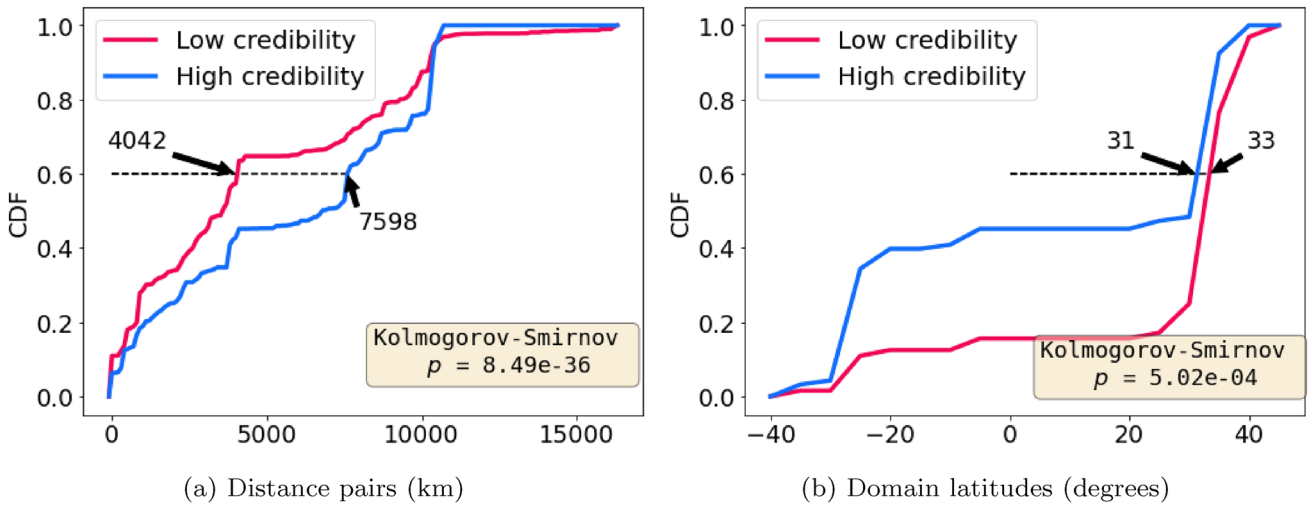


Fig. 2 CDF of geolocation numerical attributes with the lowest similarity

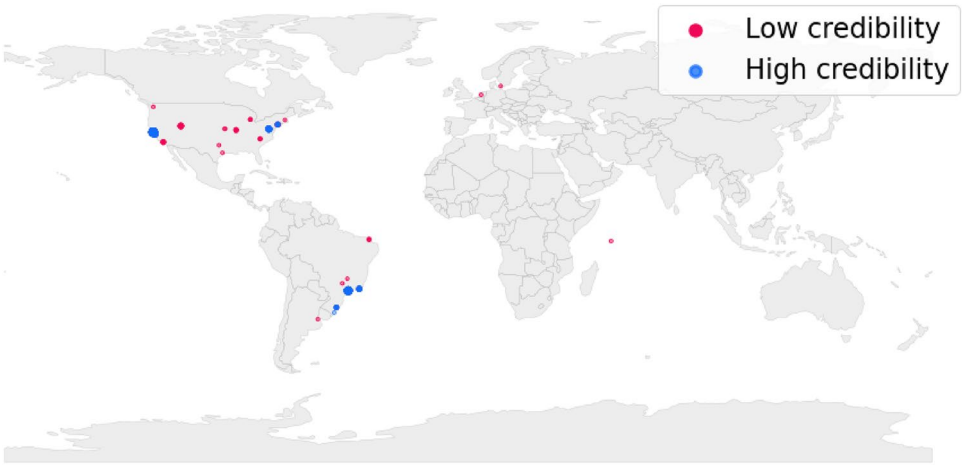


Fig. 3 Map displaying IP geolocation by credibility group. Point sizes are scaled by the number of sites in the region

Latitude Many cloud computing or online content hosting services are widely available and accessible both in Brazil and abroad. Latitude allows us to directly investigate in which hemisphere websites are hosted. In particular, most computing centers (*datacenters*) and hosting services abroad are concentrated in the United States and Europe.

Figure 2b indicates that low-credibility websites are more likely to be hosted in the northern hemisphere. Here it is worth noting that zero latitude separates the hemispheres and, therefore, through the map in Fig. 3 we can correlate latitudes greater than zero with sites registered abroad (sites hosted in the northern part of Brazil were geolocated in Fortaleza (Ceará state in Brazil) and Parauapebas-PA (Pará state in Brazil, which are below the Equator). Figure 2b confirms this notion since less than 20% of low-credibility websites are hosted below the equator, while for high-credibility sites the figure is approximately 50%.

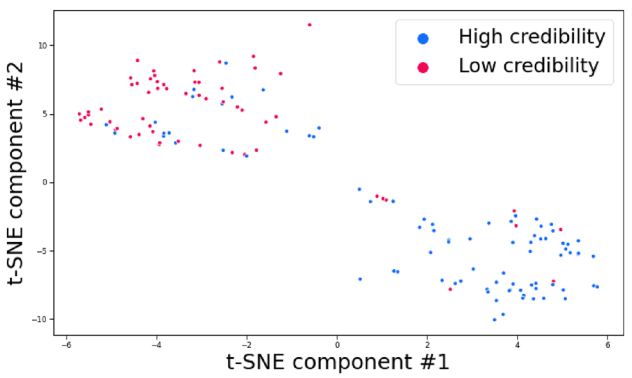


Fig. 4 Discriminative power of websites attributes using the t-SNE dimensionality-reduction technique

Table 6 Experimental setup of the binary classification

Dataset (#instances)	169 websites
Classifier	Random forest (RF)
Cross validation	5-fold
Criterion	Cross-entropy
Number of attributes	Square root of total
Criterion	Cross-entropy

Table 7 Results obtained by RF classifier w.r.t AUC-ROC, F1 Score, and Accuracy

	Test set	Validation sets
AUC-ROC	0.828	0.845 (± 0.025)
F1 Score	0.814	0.833 (± 0.034)
Accuracy	0.812	0.834 (± 0.036)

6 Usefulness of attributes

In order to investigate the real-world applicability of the all attributes explored in this work, it is important to survey their collective potential to differentiate high- and low-credibility websites. To this end, we apply a dimensionality reduction technique, namely, t-SNE (van der Maaten and Hinton 2008), to visualize the separation between the classes enabled by our attributes.

Figure 4 presents a two-dimensional map in which each data point is given a location based on its calculated attribute values. Notably, we can observe two spatially separated clusters of data points, each primarily made up by either credibility class. This suggests that the attributes explored in this work can be useful to distinguish low-credibility news websites from their high-credibility counterparts. We note that considering additional dimensions would provide even greater separation between the two sets.

7 Automated identification of low-credibility websites

In the last section, we observed that low-credibility websites in Brazil offer distinctive combinations of the characteristics presented in this work. Here, we attempt to leverage this nature to develop a based machine learning model that is capable of setting apart low-credibility websites from their high-credibility counterparts. Table 6 presents the experimental setup through which we surveyed the prediction capabilities of our attributes.

As aforementioned (see Table 1) the dataset explored in this work is composed of 71 low-credibility websites and 98 high-credibility websites, totaling 169 instances. We train an array of models through Grid Search[15] to find the best-fit combination of hyper-parameters. In this process, we experiment with different loss functions (Gini impurity, cross-entropy) as well as the maximum number of features (square root of the total or unlimited), resulting on the usage of cross-entropy and capping the limit of features used to the square root of the total. Additionally, in order to account for the slightly different frequencies between the two classes, we employ balanced class-weight during training, where each class is attributed a weight inversely proportional to its frequency. Finally, 5-fold cross-validation is employed to avoid overfitting the data, thus enhancing generalization capabilities. The ensuing model is evaluated through AUC-ROC, F1, and accuracy scores (Baeza-Yates and Ribeiro-Neto 1999). Table 7 presents these metrics.

The performance metrics obtained by the model in the test set (measured over novel test data not previously used for training or parameter fine-tuning) is 0.828, 0.814, and 0.812 for AUC-ROC, F1 Score, and Accuracy, respectively, with relatively low variability observed when measured over the validation sets of each fold (see Column “Validation Sets" in Table 7). This indicates that the classifier is capable of detecting low news credibility websites in Brazil with reasonably good performance and could be employed as a flagging system for more detailed human analysis.

8 Conclusion and future work

In this paper we present a broad characterization of Brazilian high- and low-credibility news websites, highlighting their differences from the perspective of network attributes. To allow for such analysis, we created a set of low-credibility websites by identifying websites that have published at least one news piece that had its veracity contested by an accredited fact-checking agency. We then gathered public information about those websites and we computed a total of 31 attributes including domain, certificate, and geolocation characteristics.

Our results reveal that only 12.2% of low-credibility websites are registered and hosted in Brazil against 47% for high-credibility websites. In addition, we have shown that highly credible websites have a longer lifespan, are held under domain names registered for a longer duration, and have TLS certificates that are valid for longer periods. We also found that low-credibility websites are more likely to use alternative TLDs. Finally, we demonstrated the existence of a separation between high and low-credibility websites from the perspective of our attributes by visualizing them

[15] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

in 2 dimensions via t-SNE and later training a random forest model capable of predicting the credibility of websites in our dataset with good performance. In unseen test data, the model was able to achieve an AUC-ROC of 0.828%. In short, these results identify a set of attributes with discriminative power, contributing to the understanding of the phenomenon of misinformation.

We hope that our findings reveal patterns of low-credibility websites that can be useful to distinguish them from others, thus opening a new avenue for future efforts in this field. As future work, we intend to expand our dataset to include a wider variety of low-credibility websites, including but not limited to websites whose low-credibility nature is subtler. Furthermore, we plan to explore more robust machine learning techniques better equipped to explore the separation between low and high-credibility websites in this new dataset. Particularly, we intent to expand our usage of model explainability methods which in turn will increase the interpretability of our results and allow for better reasoning about network attributes. Finally, an important evolution will be measuring the effectiveness of our contributions on a dataset pertaining to news websites active in other countries. This will allow us to observe how general our proposed computer network attributes are on study of misinformation ecosystem around the world.

Acknowledgements This work was partially supported by the Ministério Público de Minas Gerais (MPMG), project Analytical Capabilities, as well as grants from CNPq, CAPES, FAPEMIG, and FAPESP.

Author contributions JMMC prepared Figs. 1, 2 and 3, Tables 2, 3 and 4 and wrote sections 3–5 JCSR prepared Figure 4, Table 1, and wrote sections 6, 7 FB wrote sections 1, 2, 7 and conceived the initial concept. JCSR and FB reviewed the manuscript.

Data availability The data collected from Brazilian fack-checking agencies used in this study is publicly available here: https://doi.org/10.5281/zenodo.5191798. Other datasets generated during and/or analyzed on this work are obtainable from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

References

Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. J Econ Perspect 31(2):211–236

Araujo L, Couto JMM, Nery LF, Rodrigues I, Almeida J, Reis JCS, Benevenuto F (2024) Finding fake news websites in the wild. In: arXiv

Baeza-Yates R, Ribeiro-Neto B (1999) Modern information retrieval, vol 463. ACM press, New York

Bazmi P, Asadpour M, Shakery A (2023) Multi-view co-attention network for fake news detection by modeling topic-specific user and news source credibility. Inf Process Manag 60(1):103146

Bozarth L, Budak C (2020) Market forces: quantifying the role of top credible ad servers in the fake news ecosystem. In: Proceedings of the international AAAI conference on web and social media (ICWSM)

Couto JM, Reis JC, Cunha Í, Araújo L, Benevenuto F (2022) Characterizing low credibility websites in brazil through computer networking attributes. In: Proceedings of the international IEEE/ACM conference on advances in social networks analysis and mining (ASONAM) pp 42–46

Daigle L (2004) Whois protocol specification. Technical report, Network working group

Depoux A, Martin S, Karafillakis E, Preet R, Wilder-Smith A, Larson H (2020) The pandemic of social media panic travels faster than the covid-19 outbreak

Ferrara E (2017) Disinformation and social bot operations in the run up to the 2017 french presidential election. First Monday 22(8)

Ferrara E, Varol O, Davis C, Menczer F, Flammini A (2016) The rise of social bots. Commun ACM 59(7):96–104

Fletcher R, Cornia A, Graves L, Nielsen RK (2018) Measuring the reach of "fake news" and online disinformation in Europe. Technical report Reuters Institution and University of Oxford

Giełczyk A, Wawrzyniak R, Choraś M (2019) Evaluation of the existing tools for fake news detection. In: Proceedings of the IFIP international conference on computer information systems and industrial management (CISIM), pp 144–151

Helmstetter S, Paulheim H (2018) Weakly supervised learning for fake news detection on twitter. In: Proceedings of the IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), pp 274–277

Hussain MN, Tokdemir S, Agarwal N, Al-Khateeb S (2018) Analyzing disinformation and crowd manipulation tactics on youtube. In: Proceedings of the IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), pp 1092–1095

Júnior M, Melo P, Kansaon D, Mafra V, Sá K, Benevenuto F (2022) Telegram monitor: monitoring brazilian political groups and channels on telegram. In: Proceedings of the ACM conference on hypertext and social media (HYPERTEXT), pp 228–231

Lazer DM, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, Rothschild D et al (2018) The science of fake news. Science 359(6380):1094–1096

Marques I, Salles I, Couto JM, Pimenta BC, Assis S, Reis JC, da Silva APC, de Almeida JM, Benevenuto F (2022) A comprehensive dataset of Brazilian fact-checking stories. J Inf Data Manag (JIDM) 13(1):2354

Massey FJ Jr (1951) The Kolmogorov-Smirnov test for goodness of fit. J Am Stat Assoc 46(253):68–78

Melo P, Messias J, Resende G, Garimella K, Almeida J, Benevenuto F (2019) Whatsapp monitor: a fact-checking system for whatsapp. In: Proceedings of the international AAAI conference on web and social media (ICWSM)

Pérez-Rosas V, Kleinberg B, Lefevre A, Mihalcea R (2017) Automatic detection of fake news. In: Proceedings of the international conference on computational linguistics, pp 3391–3401

Rath B, Salecha A, Srivastava J (2022) Fake news spreader detection using trust-based strategies in social networks with bot filtration. Soc Netw Anal Min 12(1):66

Reis JC, Correia A, Murai F, Veloso A, Benevenuto F (2019) Supervised learning for fake news detection. IEEE Intell Syst 34(2):76–81

Resende G, Melo P, Sousa H, Messias J, Vasconcelos M, Almeida J, Benevenuto F (2019) (mis)information dissemination in whatsapp:

gathering, analyzing and countermeasures. In: Proceedings of the web conference (WWW), pp 818–828
Ribeiro MH, Ottoni R, West R, Almeida VA, Meira Jr W (2020) Auditing radicalization pathways on youtube. In: Proceedings of the conference on fairness, accountability, and transparency (FAT), pp 131–141
Saez-Trumper D (2014) Fake tweet buster: a webtool to identify users promoting fake news on twitter. In: Proceedings of the ACM conference on hypertext and social media (HYPERTEXT), pp 316–317
Shao C, Ciampaglia GL, Flammini A, Menczer F (2016) Hoaxy: a platform for tracking online misinformation. In: Proceedings of the international ACM conference on world wide web (WWW) companion, pp 745–750
Shu K, Wang S, Liu H (2018) Understanding user profiles on social media for fake news detection. In: IEEE conference on multimedia information processing and retrieval (MIPR), pp 430–435
Silva M, Oliveira LSD, Andreou A, Melo POVd, Goga O, Benevenuto F (2020) Facebook ads monitor: an independent auditing system for political ads on facebook. In: Proceedings of the web conference (WWW)
Spohr D (2017) Fake news and ideological polarization: filter bubbles and selective exposure on social media. Bus Inf Rev 34(3):150–160
van Der Linden S, Roozenbeek J, Compton J (2020) Inoculating against fake news about covid-19. Front Psychol 11:2928
van der Maaten L, Hinton G (2008) Visualizing data using t-sne. J Mach Learn Res 9(86):2579–2605
Volkova S, Shaffer K, Jang JY, Hodas N (2017) Separating facts from fiction: linguistic models to classify suspicious and trusted news posts on twitter. In: Proceedings of the annual meeting of the association for computational linguistics (ACL), pp 647–653
Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380):1146–1151