Characterizing Low Credibility Websites in Brazil through Computer Networking Attributes

João M. M. Couto*, Julio C. S. Reis†, Ítalo Cunha*, Leandro Araújo*, Fabrício Benevenuto*
*Universidade Federal de Minas Gerais (UFMG) – Brazil, †Universidade Federal de Viçosa (UFV) – Brazil
{joaocouto, cunha, leandroaraujo, fabricio}@dcc.ufmg.br, jreis@ufv.br

***Abstract*—A key gear in most misinformation ecosystems is the deployment of fake news websites that publish news in a similar fashion to how news articles are put out by credible sources. The content offered by these sites is disseminated in a complex process that may involve automation, exploitation of message apps and social network algorithms, political bias, and targeted ads to reach large and niche audiences. Due to this high complexity and the rapidly evolving nature of the problem, we are just beginning to understand patterns in the various misinformation ecosystems on the Web. In this work, we offer a first step towards understanding network properties, including data from DNS records, domain registration, TLS certificates, and hosting infrastructure of Brazilian websites associated with the dissemination of misinformation content on digital platforms. Our findings, in addition to providing a better understanding of the misinformation ecosystem in Brazil, also reveal a novel set of features useful to distinguish low credibility websites from others.**

***Index Terms*—Misinformation, Fake news, Credibility, Brazilian Websites, Computer Network Attributes**

I. Introduction

Misinformation has affected several countries in recent years, endangering the integrity of public discourse, electoral processes, and democratic governance [1]–[3]. In the Coronavirus pandemic, for instance, the issue has reached new levels, now including the diminishment of public health concerns, promotion of medication without proven efficacy, or dismissal of sanitary measures [4], [5]. Authorities have recognized the problem throughout the world: 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 [6].

Particularly in Brazil, the 2018 presidential elections were 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 [7]–[9]. In this context, misinformation has gained an unprecedented magnitude in the country, greatly empowered by digital platforms. The observed effectiveness of those campaigns now brings about widespread concern that this phenomenon will recount itself in the 2022 presidential elections.

Misinformation campaigns are increasingly becoming complex, not only exploiting a whole array of digital platforms for misinformation dissemination, but also exploring different techniques to artificially maximize reach [10]–[12]. A key component of any 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 address the problem of measuring and understanding characteristics of this kind of websites. Particularly, we focus on characterizing networking attributes, which, to the best of our knowledge, are not explored in previous studies.

To do that, we build a large 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. Then, we gathered a series of 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 use this information to extract attributes and characterize these websites dedicated to disseminating misinformation 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.

II. Dataset Construction

Ideally, we would like to have at our disposal a curated list of low and high credibility Brazilian websites. However, in Brazil, a list of low credibility 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 [13]. Thus, unfortunately, such a list is not available, compelling us to construct one from scratch. In this section, we describe our strategy to build a dataset of low and high credibility websites.

A. 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.

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00 É 2022 IEEE

First, we identify a initial "seed" consisting of a news article containing misinformation. Then, we collect all tweets made by users who have posted a link to this article on Twitter. We then rank the websites associated with those links by their H-index and consider the top-k most tweeted links of each. For each top link, we search, through keywords and a fact-check repository [14] available at https://zenodo.org/record/5191798, for a fact-check debunking it's claims. Articles successfully matched are rerun as seeds and their associated websites are added to our low credibility websites dataset. More details about our proposed obtained can be found in [15].

B. High credibility websites

In this work, we consider high credibility websites all websites accredited by Brazil's National Association of Newspapers – anj.org.br (ANJ), a nonprofit organization recognized internationally for observing the principles of responsibility, especially in the context of misinformation prevention. ANJ's statute observes mandatory compliance to their code of ethics, which includes the verification of the truth of published facts.

III. Attribute Categories

The deployment and maintenance of a website on the Internet requires 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 have the potential to allow monitoring systems to quickly flag low credibility websites.

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

A. Domain attributes

The resolution of domain names via the Domain Name System (DNS) is essential for public websites. A domain's registration and the configuration of its authoritative DNS servers provide information about its operation. Moreover, they enable the identification of patterns that serve as indication of the robustness of a given website's infrastructure and possibly the registrant's intentions when the domain was created.

An example is the total duration a website's registrant purchases a domain name for. At the start of a new misinformation campaign, registrants are aware that newly registered domains are likely to be taken down by court order. Therefore, misinformation websites tends to observe shorter registration duration to minimize financial losses. In our work, we implemented 14 attributes in the *domain* category. These attributes are associated with the registration and configuration of the domain name, including DNS data.

B. Certificate Attributes

Encrypted access to public websites is an increasingly adopted practice on the Internet, mainly through the use of the *HyperText Transfer Protocol Secure* (HTTPS).

Offering HTTPS on a website involves issuing TLS certificates necessary to authenticate the identity of servers responding to requests directed to a domain. TLS certificate generation and maintenance services often incur recurring costs related to the level support, robustness, or flexibility of certificate properties. Free-tier TLS certification services generally issue certificates valid for less than or equal to 90 days versus 1+ years for paid ones. As an example, our results show that only a tiny portion of low credibility websites employ long-lasting certificates while a quarter of high credibility websites do. In total, we implemented 10 *certificate* attributes related to the TLS certificate (or lack thereof) used by a website.

C. Geolocalization attributes

In Brazil, recurrent government initiatives aim to eliminate misinformation vectors [16].Initiatives of this nature have 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. Thus, the use of domain registration and hosting services offered by foreign entities allow websites to operate under the jurisdiction of another country. Through this, 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 contains useful information to help differentiate low credibility news sources from their high credibility counterparts. We implemented 7 *geolocation* attributes derived from the IP address of websites and their associated Autonomous Systems.

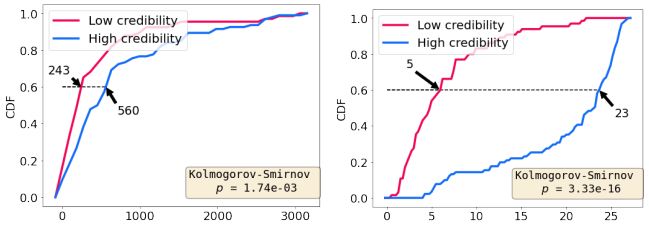
IV. Results

In this section, we present the results for the three sets of attributes presented in §III. Our aim is to characterize both credibility groups and investigate the discriminative capacity of the extracted attributes. We demonstrate that the attributes calculated on the set of low credibility websites follow fundamentally different distributions from websites in the high credibility set. We report the mean, 60^{th} percentile, and the p-value of the Kolmogorov-Smirnov (KS) test [17] for the two numerical attributes with the most dissimilar distributions in each attribute category. The KS test computes 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 websites, *i.e.,* they can serve as input attributes or features to a classifier model.

For categorical and Boolean attributes, we present the incidence, in percentage points, of each category (false or true

TABLE I: Extracted attributes aggregated by category and data type (i.e., Bool, Num. = Numerical, Cat. = Categorical).

Category	Type	Attributes list
Domain	Bool	Subdomain contains a hyphen ‖ Subdomain contains a digit ‖ TLD (Top Level Domain) is either .br or .com (most common TLDs in the dataset) ‖ URL contains a journalistic keyword (e.g., gazette, tribune) ‖ Registrant enabled WHOIS privacy options
	Num.	Number of hops necessary to resolve subdomian into a IP address ‖ Number of CAA or TXT entires in the domain's DNS ‖ Number of characters in the subdomain ‖ Time, in days, since the initial registration of the domain ‖ Time, in days, until the domain expiry date ‖ Time, in days, since the domain's DNS was last modified
	Cat.	Autonomous System Number associated with the subdomain's IP ‖ Registrar utilized for registering the domain (e.g., GoDaddy) ‖ Registrar URL (e.g., NameCheap.com)
Certificate	Bool	Domain is acessible via HTTP requests ‖ Server redirects requests HTTP ‖ Certificate Issuer is the popular "Let's Encrypt" free certification service ‖ TLS certificate has expired
	Num.	Number of bits present in the domain's public key used for TLS handshake ‖ Time since the TLS certificate was issued (days) ‖ Time until the expiry date of the TLS certificate (days) ‖ Total TLS certificate lifespan (issuing until expiry)
	Cat.	TLS certificate issuing entity ‖ Country code associated with the TLS certificate issuing entity
Geolocation	Bool	Geolocation of IP address resulted in coordinates within Brazil ‖ Geolocation of IP address resulted in coordinates within the U.S. ‖ IP address coordinates within the country associated with ASN
	Num.	Geolocated IP address latitude ‖ Geolocated IP address longitude
	Cat.	Country code associated with the IP address coordinates ‖ Country code associated with the ASN



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

Fig. 1: Comparison of the two most dissimilar numerical attributes between high- and low credibility websites (in days).

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 websites.

A. Domain attributes

A domain name's registration attributes (*e.g.,* registrant, age, expiration date) can be obtained by querying WHOIS servers. In addition, we obtain configuration attribute values regarding the domain's authoritative server (*e.g.,* number of CAA/TXT entries) performing DNS queries using the dig tool. In both cases, query results are filtered to extract fields of interest using Python scripts that parse 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 websites pay registration services for a shorter duration when compared to high credibility websites. Figure 1a corroborates that possibility: 60% of low credibility websites expire within 243 days while the same value for high credibility websites 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 identification 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 maintain and create 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 websites are less than 5 years old. Websites within the high credibility set have a much longer lifespan: only 6.7% 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 II we can observe a large discrepancy for the country of IP geolocation: only 13.6% of low credibility websites are hosted in Brazil, while the incidence among high credibility websites is 45.8%. This result indicates that low credibility websites are operated abroad significantly more often than high credibility websites. We find similar results for the Autonomous System (ASN) country of registration. Most TLS certificates are issued by American companies for both classes of websites.

Among the Boolean attributes presented in Table III, 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 (news) 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).

B. Certificate attributes

To evaluate our intuition that the credibility of websites can be captured by the TLS certificate attributes proposed in §III-B, we extract certificate attributes via commands offered by the OpenSSL library and curl. We used OpenSSL to extract attributes associated with the certificates themselves, such as the issuance and expiration dates, and curl to determine

TABLE II: Geolocation for categorical attributes by credibility group (LC = Low Credibility, HC = High credibility). "Others" includes CN, SC, AR, NL, and 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

TABLE III: Occurrence of true label per credibility group and boolean attribute.

		Low Credibility(%)	**High Credibility (%)**
D	subdomain-hifen	9.9	2.0
	subdomain-digit	15.5	6.1
	news-keywords	22.5	49.0
	tld-br-or-com	91.5	98.0
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

TABLE IV: Distribution of certificate lifespans (days)

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

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. We investigated whether certificates issued to high credibility websites have a longer validity period than those issued to low credibility 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, and thus summarize the results in incidence Table IV. While 23.9% of certificates from highly credible websites last for more than one year, this is the case for only 2.4% of low credibility websites. This result indicates that certificates with a duration of more than one year are 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 month [18].

Categorical and Boolean certificate attributes. The C section of Table III 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.

C. Geolocation

As aforementioned, we use the IPStack geolocation API to obtain the geographic coordinates associated with the IP address each website's domain name resolved to. Figure 2 shows the location of websites, where color encodes credibility and point sizes capture the number of websites in the region. The map intuitively suggests the correlation between the coordinates obtained and the presence of misinformation content.

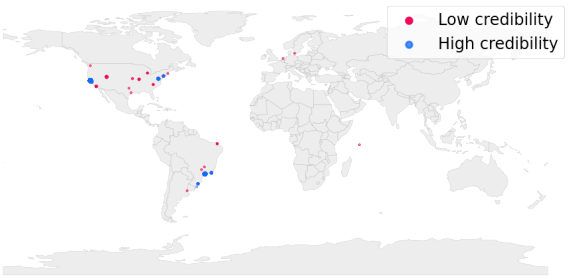


Fig. 2: Map displaying IP geolocation by credibility group. Point sizes scaled by the number of websites in the region.

Geographical concentration. Many of the main hosting services are located in major technological hubs. Figure 2 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 websites are spread in 28 different cities, suggesting that low credibility websites are less concentrated in large hubs. High credibility websites are concentrated in Sao Paulo and the Bay Area, the two major tech hubs in Brazil and the US.

Latitude. Many cloud computing or online content hosting services are widely available and accessible both in Brazil and abroad. We find that less than 20% of low credibility websites are hosted below the Equator, while for high credibility websites the figure is approximately 50%. Low credibility websites are more likely to be hosted in the northern hemisphere as they are more often hosted outside of Brazil and hosting services abroad are concentrated in the United States and Europe.

V. Conclusion

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 findings reveal properties of low credibility websites that can be useful to distinguish them from high credibility ones, thus opening a new avenue for future efforts that can be explored by the misinformation research community. We plan to explore machine learning and other techniques to automatically identify low credibility websites in the wild.

Acknowledgments. This work was partially supported by grants from MPMG, project Analytical Capabilities, FAPEMIG, FAPESP, CNPq, and CAPES.

References

[1] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

[2] D. Spohr, "Fake news and ideological polarization: Filter bubbles and selective exposure on social media," *Business Information Review*, vol. 34, no. 3, pp. 150–160, 2017.

[3] E. Ferrara, "Disinformation and social bot operations in the run up to the 2017 french presidential election," *First Monday*, vol. 22, no. 8, 2017.

[4] S. van Der Linden, J. Roozenbeek, and J. Compton, "Inoculating against fake news about covid-19," *Frontiers in psychology*, vol. 11, p. 2928, 2020.

[5] A. Depoux, S. Martin, E. Karafillakis, R. Preet, A. Wilder-Smith, and H. Larson, "The pandemic of social media panic travels faster than the covid-19 outbreak," 2020.

[6] "The New York Times. Nobel Peace Prize Awarded to 2 Journalists, Highlighting Fight for Press Freedom," https://www.nytimes.com/live/2021/10/08/world/nobel-prize, October 2021.

[7] C. Tardaguila, F. Benevenuto, and P. Ortellado, "The new york times. fake news is poisoning brazilian politics. whatsapp can stop it," https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html, October 2018.

[8] G. Resende, P. Melo, H. Sousa, J. Messias, M. Vasconcelos, J. Almeida, and F. Benevenuto, "(mis)information dissemination in whatsapp: Gathering, analyzing and countermeasures," in *Proc. of The Web Conference (WWW)*, 2019, pp. 818–828.

[9] P. Melo, J. Messias, G. Resende, K. Garimella, J. Almeida, and F. Benevenuto, "Whatsapp monitor: A fact-checking system for whatsapp," in *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, 2019, pp. 676–677.

[10] M. Silva, L. S. d. Oliveira, A. Andreou, P. O. V. d. Melo, O. Goga, and F. Benevenuto, "Facebook ads monitor: An independent auditing system for political ads on facebook," 2020, pp. 224–234.

[11] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Communications of the ACM*, vol. 59, no. 7, pp. 96–104, 2016.

[12] M. H. Ribeiro, R. Ottoni, R. West, V. A. Almeida, and W. Meira Jr, "Auditing radicalization pathways on youtube," in *Proc. of the Conference on Fairness, Accountability, and Transparency (FAT)*, 2020, pp. 131–141.

[13] "Revista Oeste. Agência de checagem Aos Fatos é condenada por publicar fake news (in Portuguese)," https://revistaoeste.com/brasil/agencia-de-checagem-aos-fatos-e-condenada-por-publicar-fake-news/, May 2022.

[14] J. Couto, B. Pimenta, I. M. de Araújo, S. Assis, J. C. S. Reis, A. P. da Silva, J. Almeida, and F. Benevenuto, "Central de fatos: Um repositório de checagens de fatos (in portuguese)," in *Proc. of the Brazilian Symposium on Databases (SBBD) - Dataset Showcase*, 2021, pp. 128–137.

[15] L. Araújo, L. F. Nery, I. C. Rodrigues, J. M. Couto, J. Reis, A. P. C. Silva, J. Almeida, and F. Benevenuto, "Identificando websites de desinformação no brasil (in portuguese)," in *Proc. of the Brazilian Symposium on Databases (SBBD)*, 2022, pp. 355–360.

[16] "Metropolitano Manaus News. Justiça Eleitoral derruba nove fake news lançadas contra David Almeida (in Portuguese)," https://metropolitanomanaus.news/justica-eleitoral-derruba-nove-fake-news-lancadas-contra-david-almeida/, October 2022.

[17] F. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[18] J. Aas, "Let's encrypt. why ninety-day lifetimes for certificates?" https://letsencrypt.org/2015/11/09/why-90-days.html, November 2015.