


Automatic Inference of Brazilian Websites' Reliability for Combating Fake News: Domain and Geolocation Features

Marcos Paulo Cezar de Mendonça   [Universidade Federal Fluminense | marcos_cezar@ic.uff.br]

Igor Monteiro Moraes  [Universidade Federal Fluminense | igor@ic.uff.br]

Diogo Menezes Ferrazani Mattos  [Universidade Federal Fluminense | menezes@midiacon.uff.br]

 Institute of Computing, Universidade Federal Fluminense, Av. Gal. Milton Tavares de Souza, s/n, São Domingos, Niterói, RJ, 24210-590, Brazil.

Received: 04 October 2024 • **Accepted:** 03 March 2025 • **Published:** 19 March 2025

Abstract Evaluating the reliability of websites that propagate news is critical in combating disinformation. Websites with low reliability often serve as the breeding ground for fake news that spreads rapidly across social networks. In response, this paper introduces an automatic evaluation approach to assessing the reliability of Brazilian websites by analyzing network-related features, eliminating the need for exhaustive content scanning. Unlike previous methodologies focused on social network analysis, our approach leverages publicly available website features, including domain-related features, geolocation data, and TLS certificate attributes. The paper proposes a supervised learning model and curates a comprehensive dataset comprising reliable and unreliable sites. Through rigorous training and evaluation using disjoint data, the model achieves an accuracy greater than 75%, effectively pinpointing reliable content websites.

Keywords: Data Analysis, Machine Learning, Classification, Fake News, Websites Reliability

1 Introduction

The spread of fake news is a significant research issue that follows a pattern distinct from the spread of reliable news. Fake news includes persuasive headlines and copy design elements to attract readers' attention [Posetti and Matthews, 2018]. Informally, the terms fake news and disinformation are commonly used as synonyms. However, these concepts are formally different. Wardle and Derakhshan define Fake News as a wide range of disinformation, misinformation, and malinformation. Disinformation is false information that is intentionally intended to cause harm and confusion; as opposed to misinformation, it is erroneous information spread without the aim of causing harm. Misinformation consists of disseminating correct facts that are out of context and cause harm. When comparing fake news and disinformation, we highlight that fake news intentionally deceives the reader. In contrast, disinformation results from the dissemination of false information in different media to deceive and manipulate the public [Wardle and Derakhshan, 2017; de Oliveira *et al.*, 2021].

When analyzing the situation in Brazil regarding the effects of disinformation and fake news, we can observe a significant challenge in the political realm, particularly in public discourse and electoral contests. This problem has also had far-reaching consequences in other areas, such as public health, with the spread of fake news about the Coronavirus pandemic. The Brazilian presidential elections in 2018 were rife with disinformation and manipulation of facts through digital channels, particularly on messaging platforms such as WhatsApp. The characteristic of rapid dissemination of information to a large audience without requiring prior content analysis made it easier for these false messages to spread

rapidly. In the 2022 elections, the situation repeated, but the Telegram application played a major role in political mobilization efforts [Ramos *et al.*, 2022; Júnior *et al.*, 2021].

To mitigate and minimize the impacts of disinformation, the Strategic Plan for the 2022 Elections, developed by the Brazilian Superior Electoral Court (*Tribunal Superior Eleitoral* - TSE), established a permanent program to confront disinformation. This program considers any content that may be 'potential disinformation' to include material that is deemed 'false, mistaken, deceptive, inaccurate, manipulated, fabricated, fraudulent, unlawful, or hateful, as well as content that is taken out of context in any format, dissemination channel, or by any intention of the agent. The program includes educational initiatives to inform, empower, and respond to the public on the electoral process [BRASIL, 2022].

On the other hand, the vast amount of information generated daily makes manual classification of the reliability of a website¹ a burdensome and unfeasible task, which requires the help of computational strategies [Hua *et al.*, 2023; Medeiros *et al.*, 2020]. Many studies currently validate the ability of machine learning models to detect malicious bias in sites, for example, phishing, and identify spreaders of fake news [Alkawaz *et al.*, 2021]. However, most strategies analyze the website's content to classify the information's reliability.

This paper proposes classifying websites into reliable and unreliable categories using supervised machine learning models such as Naive Bayes, Support Vector Machine, Decision Tree, Random Forest, Neural Networks, Linear Regression, and K-Nearest Neighbors. The models will use web-

¹In this paper, we define website reliability as the measurement of the correctness and the reliability of the content disseminated by the website.

site features such as domain, certificate, and geolocation information to classify. The proposed approach differs from previous works, which primarily focused on on-site content for detection. Instead, our proposal emphasizes the analysis of a website's geographic and domain features. These publicly available features allow for inference of site reliability with low computational resource commitment. The results demonstrate the effectiveness of our approach, with an accuracy of over 75%.

The paper is organized as follows. Section 2 lists related works. Section 3 presents our proposal. Section 4 describes the data collection methodology and dataset construction. Section 5 evaluates the proposal and highlights the main findings. Section 6 concludes the paper.

2 Related Work

The growing interest in combating the impacts of fake news has motivated researchers, particularly in computer science, to propose effective alternatives against its spread. One approach to analyzing disinformation and fake news dissemination in Brazil is the 'FakeSpread' framework, which utilizes graph theory to track the spread from a source [Cordeiro *et al.*, 2020]. The work is based on websites reported in investigation of the Brazilian Mixed Parliamentary Inquiry Committee (*Comissão Parlamentar Mista de Inquérito - CPMI*) on Fake News *Fake News*² to identify other sites that use them as a source of disinformation and, thus, establish a relationship between the sites and the sources. Although the work neglected to classify the reliability of a website, it focused on studying the propagation of fake news and contributed to finding a relationship between an unreliable website and unreliable sources.

Another approach, conducted by Couto *et al.*, categorizes websites into low and high credibility based on data collected from X (formerly Twitter), identifying users who posted the news as the root source [Couto *et al.*, 2022]. The authors collect all the news previously posted by these users. From these sources, certification, registration, and location features are obtained. The work does not employ machine learning models for automating classification but allows for inferring meaningful relationships between low and high-credibility websites by comparing their features.

Additionally, Baly *et al.* have used features extracted from website URLs to classify reliable and unreliable websites. Their study aimed to evaluate political bias and the reliability of information sources by analyzing samples of articles published on Wikipedia and X (formerly Twitter), along with their traffic and URL structure. This analysis was conducted using machine learning algorithms [Baly *et al.*, 2018]. These studies reinforce the utility of features derived from the URL structure in validating the reliability of information.

Reis *et al.* utilized machine learning models to identify fake news in news articles about the 2016 US elections. The models took into account features such as political bias and domain location. The data collected was classified based

upon the key features of low-reliability sites found in the literature. Additionally, the researchers proposed a new set of features, including political bias, reliability, and domain location [Reis *et al.*, 2019].

Saleem Raja *et al.* and Ahammad *et al.* focus on the classification of phishing websites using lexical and Whois registration features, proposing the inclusion of features such as SSL/TLS certificate attributes and geolocation [Saleem Raja *et al.*, 2021; Ahammad *et al.*, 2022]. Xuan *et al.* and Palaniappan *et al.* also explore lexical features of URLs to classify malicious sites, employing machine learning models such as Random Forest and Logistic Regression, respectively [Xuan *et al.*, 2020; Palaniappan *et al.*, 2020].

Previous work, including graph analysis, URL lexical features, and machine learning models, have been used to classify malicious websites and detect fake news. However, many of these studies focused on content classification based on social media data or detailed text analysis of news articles. In contrast, the present proposal distinguishes by automatically evaluating website reliability and identifying the source of fake news directly from domain features, geolocation, and SSL/TLS certificate attributes. This approach eliminates the need to scan the entire website's content, providing a more efficient and straightforward approach to mitigating the spread of online disinformation.

In this context, these works establish correlations or leverage attributes such as domain characteristics, certifications, and geolocation to evaluate the reliability of websites. By distinguishing between reliable and unreliable sources, they provide a framework for reliability assessment. This approach employs machine learning algorithms to analyze these attributes and determine, based on patterns observed in the data, the likelihood of a website being reliable or unreliable.

However, it is important to note that these relationships are not causal but represent statistical associations derived from the dataset. The identified attributes do not serve as deterministic reliability indicators but offer logical implications grounded in the trends uncovered through analysis.

3 Proposal for Website Reliability Classification

This work applies supervised machine learning algorithms to classify reliable and unreliable news websites considering network features. Supervised machine learning algorithms use pre-categorized data and its features for training and pattern learning. These patterns are then employed for category classification or value prediction [Sen *et al.*, 2020].

A dataset must be used to train the algorithms using examples to classify with machine learning algorithms. This dataset should contain features describing the category of each record. When considering features that best represent the characteristics of websites to classify them as reliable or unreliable, many works use features extracted from website content analysis or their relationship with social networks. In this work, we follow the methodology of Couto *et al.*, which aims to characterize the relationship between the reliability of Brazilian news websites and features divided into three categories: domain, certificate, and geolocation [Couto *et al.*,

²Available at <https://legis.senado.leg.br/comissoes/comissao?codcol=2292>.

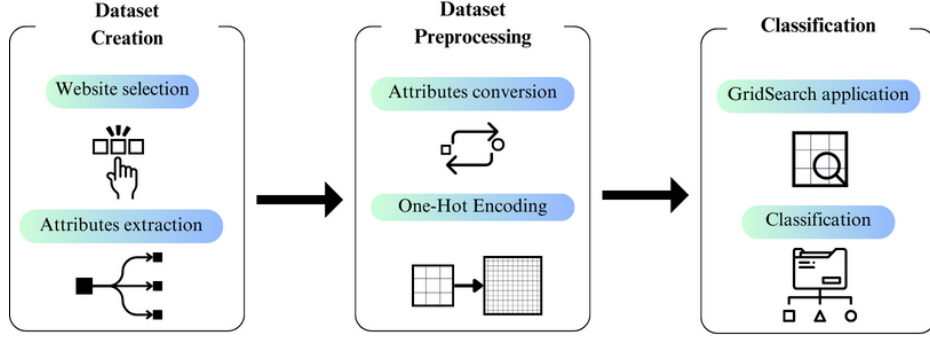


Figure 1. The website reliability classification process includes creating a dataset of websites and features, preprocessing the data, and classifying them.

2022]. These features have lower computational costs, requiring fewer resources and less time to obtain.

The domain features (Table 1) are related to the Domain Name System (DNS), which is responsible for mapping a Uniform Resource Locator (URL) to an Internet Protocol (IP) address. During the domain registration process, various information is required for its configuration. Choosing a hosting service and its servers is also necessary to generate meaningful information that may be used to characterize that website. These features contain information about the subdomain, creation, expiration, and update dates of the domain, as well as the identity of the domain registrant.

Another essential feature of this category is the Autonomous System Number (ASN). It refers to an Autonomous System operated by an organization that hosts the website. These Autonomous Systems form the network layer on the Internet and are responsible for routing between other Autonomous Systems and managing their IP addresses. Data related to the URL structure is also used, such as the size of the subdomain name, the Top-Level Domain (TLD), and keywords related to the news scope.

SSL/TLS certificate attributes (Table 2) are an integral part of website security features. They add the security layer to the HTTPS protocol and ensure the encryption of sensitive data transmitted between the client and server. SSL/TLS certificates can be obtained for any domain and contain important information such as who requested them, their creation and expiration dates, and the entity responsible for generating them, reassuring the data protection measures in place.

Geolocation features (Table 3) are also related to the physical location of a website's servers. These features describe the country and region where the servers associated with the sites are located and the geolocation information associated with the ASN.

In addition to these features listed by Couto *et al.*, this work added other ones to provide machine learning models with greater context about the websites. The additional features are: an feature that checks if the website, when accessed, redirects to the HTTPS port, the type of encryption cipher used by the SSL/TLS certificate, the region of the country of the IP address, and the ASN. Furthermore, this work also aims to study the classification of reliable and unreliable websites using network features that do not depend on the website's lexical structure.

4 Evaluation Methodology

We apply the methodology outlined in Figure 1 to evaluate our proposal. The evaluation methodology consists of three stages. Firstly, we collect URLs from both reliable and unreliable websites. Next, we extract features for each website. In the second stage, the dataset is preprocessed to prepare the data for machine learning models. Finally, in the third and final stage, we select hyperparameters for each model and carry out the classification.

4.1 Dataset Creation

To create a dataset for classifications³, we use a two-step process. First, we select reliable and unreliable news websites. Then, we extract network features from these websites to ensure that our dataset contains high-quality and reliable data.

The websites considered as a reliable source of information were selected from the National Newspaper Association (*Associação Nacional de Jornais - ANJ*)⁴. This Brazilian association advocates for the interests of newspapers and promotes their growth by sharing experiences, disseminating innovation, and fostering cooperation among similar organizations. In total, 95 websites were gathered, but five were offline during the research.

We deploy the methodology defined by Cordeiro *et al.* to obtain a potential list of unreliable websites. In their work, the authors studied how unreliable websites reference other unreliable websites [Cordeiro *et al.*, 2020]. The websites reported in the Brazilian Mixed Parliamentary Inquiry Committee (*Comissão Parlamentar Mista de Inquérito - CPMI*) investigation on Fake News are the source of a search for more websites that reference them.

We deploy the Custom Search JSON API from Google to provide a customized search engine capable of conducting searches through HTTP requests. For each website in the list of the CPMI, we searched for the first 100 news websites that cited one of the sites identified by the CPMI on Fake News as a source, if available.

We obtained 240 websites, and then a filtering process removed domains that belonged to the scope of news portals, such as social networks, duplicate references, or references to files, e.g., PDFs. Additionally, sites that were unavailable

³The authors can provide access to the dataset upon request by e-mail.

⁴Available at <https://www.anj.org.br/>.

Table 1. Table adapted from Couto *et al.* for domain attributes used in machine learning algorithms Couto *et al.* [2022].

Attribute	Data Type	Attribute Type
Hyphen in the URL	Boolean	Domain
Digit in the URL	Boolean	Domain
Top-level domain is ".br" or ".com"	Boolean	Domain
URL contains news-related keywords	Boolean	Domain
Domain registrant uses privacy for Whois queries	Boolean	Domain
Number of hops required to resolve the subdomain into an IP address	Numeric	Domain
Number of "TXT" and "CAA" entries in the DNS record	Numeric	Domain
Number of characters in the subdomain	Numeric	Domain
Days since the domain creation	Numeric	Domain
Days until the domain expiration	Numeric	Domain
Days since the last domain update	Numeric	Domain
ASN number of the website	Categorical	Domain
Name of the entity responsible for registering the domain	Categorical	Domain
URL of the entity responsible for registering the domain	Categorical	Domain

Table 2. Table adapted from Couto *et al.* for certificate attributes used in machine learning algorithms Couto *et al.* [2022].

Attribute	Data Type	Attribute Type
Server returns data from HTTP requests	Boolean	Certificate
Redirects requests made to the website	Boolean	Certificate
Redirects requests from HTTP to HTTPS*	Boolean	Certificate
The SSL/TLS certificate was issued by the "Let's Encrypt" service	Boolean	Certificate
Number of bits in the public key during the SSL/TLS handshake	Numeric	Certificate
Encryption cipher*	Categorical	Certificate
Checks if the certificate is expired	Boolean	Certificate
Days since the certificate issuance	Numeric	Certificate
Days until the certificate expiration	Numeric	Certificate
Interval of days between certificate creation and expiration	Numeric	Certificate
Entity issuing the SSL/TLS certificate	Categorical	Certificate
Country of the SSL/TLS certificate issuing entity	Categorical	Certificate

(*) Attributes added in this work.

Table 3. Table adapted from Couto *et al.* for geolocation attributes used in machine learning algorithms Couto *et al.* [2022].

Attribute	Data Type	Attribute Type
Country code of the IP address	Categorical	Geolocation
Region code of the IP address*	Categorical	Geolocation
Country code of the ASN	Categorical	Geolocation
Continent of the IP address*	Categorical	Geolocation
Continent of the ASN*	Categorical	Geolocation
Geolocation of the IP address is in Brazil	Categorical	Geolocation
Geolocation of the IP address is in the United States	Categorical	Geolocation
Latitude of the IP address	Categorical	Geolocation
Longitude of the IP address	Categorical	Geolocation

(*) Attributes added in this work.

at the time of the research were removed, resulting in a total of 132 websites labeled as unreliable.

From the selected websites, extracting features from the categories described in Section 3 was possible. We also analyzed the website's URL structure to obtain domain features, such as the presence of hyphens in the domain or the size of the subdomain.

The Python Whois library⁵ provides data on domain owners. HTTP requests to the Ninja API⁶ with a free access key retrieve information about DNS records. These records include the number of TXT and CAA entries. The number of hops reaching the website IP was obtained using the `tracert` command. We used the IPWHOIS.IO API⁷ to obtain information related to ASN data and server geolocations.

Besides, native Python libraries, such as SSL⁸ and Socket⁹, were deployed to retrieve the certificate attributes.

4.2 Dataset Preprocessing

The first part of the preprocessing involves transforming all data that were previously saved as boolean values in Python (*True* or *False*) into numerical values (1 or 0) using the Pandas library¹⁰. In the second part of the preprocessing, the *One-Hot Encode* method was employed to handle categorical data. This method transforms each distinct value of a categorical feature into a column and assigns the value 1 if the record presents that value [Al-Shehari and Alsowail, 2021].

Therefore, all categorical features were identified, which in the scope of this work are all columns containing non-

⁵Available at <https://github.com/richardpenman/whois>.

⁶Available at <https://api-ninjas.com/api/dnslookup>.

⁷Available at <https://ipwhois.io/>.

⁸Available at <https://docs.python.org/3/library/ssl.html>.

⁹Available at <https://docs.python.org/3/library/socket.html>.

¹⁰Available at <https://pandas.pydata.org/>.

numeric data and the ASN column. Although ASN is a number, it does not represent a quantity, and thus, we performed *One-Hot Encode* over ASN to avoid any attempt by any model to assign order or magnitude to these values. We carried out this process using the *OneHotEncoder* module from Scikit-learn¹¹.

4.3 Websites Classification

After preprocessing the dataset, the next step is to train and test each machine learning model to evaluate the results. In this stage, we decided to standardize and normalize the data for each model, resulting in two sets of data.

The Pipeline module from the Scikit-learn library¹² was employed to carry out standardization or normalization and also train and test the models. This module allows defining a sequence of operations to be performed. The Pipeline returns an object used to train standardized and normalized models. After building the Pipeline, we performed ten rounds of the experiment to obtain the results with a confidence interval of 95%. In each round, the dataset was split into two sets: a training set, with 80% of the original data, and a test set, comprising the remaining 20%.

After defining a set of hyperparameter values, we applied the grid search method. Ten other rounds of experiments were conducted, in which the training data was split into 90% for training within the *gridsearch* and 10% for validation. The model with a possible combination of hyperparameters was trained and evaluated in each round. The hyperparameters with the best average Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) were selected. The model was trained and tested in each round with the defined hyperparameters.

During the testing phase, the model was given the test data without any classification, and it classified each record based on the training data. Using the results of this classification and the actual classifications of the test set, a confusion matrix was created, and the model's performance metrics were evaluated and extracted.

5 Results

The performance evaluation of each machine learning model to detect reliable and unreliable websites takes into account the domain, certificate, and geolocation features extracted from each website's URL. It allows for an assessment of which model best adapts to the proposal. Additionally, metrics were calculated for each model using all the features defined in the proposal, considering only each site's network features and removing the URL's lexical features from the dataset. We also conducted an experiment to verify the relevance of the features for classification and to compare using only those presented in the work of Couto *et al.* [Couto *et al.*, 2022]. Furthermore, an investigation was carried out to characterize the most important features of reliable and

unreliable sites and understand how they are distributed in these subgroups.

The average accuracy, precision, recall, and F1-score are compared across models, considering a 95% confidence interval to evaluate the models that best fit the dataset. The Figure 2 presents the results.

According to Figure 2(a), the accuracy of all models, except for the Naive Bayes algorithm, is above 70% for the methods that consider all features, with particular emphasis on the Random Forest model, which achieved average accuracies of 77% and 76% with normalization and standardization, respectively. The models trained only with network features mostly had average accuracies below 70%.

This same behavior is observed when analyzing the results obtained for precision, shown in Figure 2(b), and the F1-score in Figure 2(d). These results indicate that the model demonstrating the best performance on the considered dataset was the Random Forest with all features. It is due to its ability to construct multiple decision trees, giving more relevance to features that reduce logarithmic loss, thus allowing for identifying more relevant features and assigning greater weights.

Other models, such as logistic regression and multilayer perceptron, achieved results comparable to Random Forest using fully normalized data within the 95% confidence interval. While these models performed slightly worse on average, their overlapping confidence intervals suggest statistically equivalent performance. Logistic regression produced average metrics of 0.79 for precision, 0.72 for accuracy, 0.73 for recall, and 0.72 for the F1-Score.

The results obtained by Naive Bayes are explained by considering that the dataset is sparse, featuring over 200 features, with some features not converging towards high decision probabilities, as there are features related to the classes that appear in the training set but not in the test set.

Unlike the previous results, Naive Bayes was the model that achieved the highest sensitivity, as seen in Figure 2(c). Thus, this model was able to have a more significant number of correct predictions considering only the positive class. This result is justified because the dataset's structure has a greater variety of values for unreliable websites than for reliable ones, allowing for a better relationship between the data in the training and test sets.

The Naive Bayes model demonstrated recall values that were statistically distinct from all other models. This distinction suggests that Naive Bayes could be instrumental in applications focused on identifying trustworthy websites where recall is the primary consideration.

When comparing the results obtained with all the features from the proposal and the network features, it is evident that the results using all the features achieved better classification performance. Thus, the URL's lexical features provide the models with greater context about the website's data, resulting in improved classification capability.

The values obtained in this study corroborate previous works [Reis *et al.*, 2019; Saleem Raja *et al.*, 2021]. Reis *et al.* highlights that the Naive Bayes model produced the worst results compared to other algorithms, while the Random Forest achieved the best results [Reis *et al.*, 2019]. Saleem Raja *et al.* employed features extracted from the URLs and do-

¹¹Available at <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html>.

¹²Available at <https://scikit-learn.org/stable/modules/compose.html>.

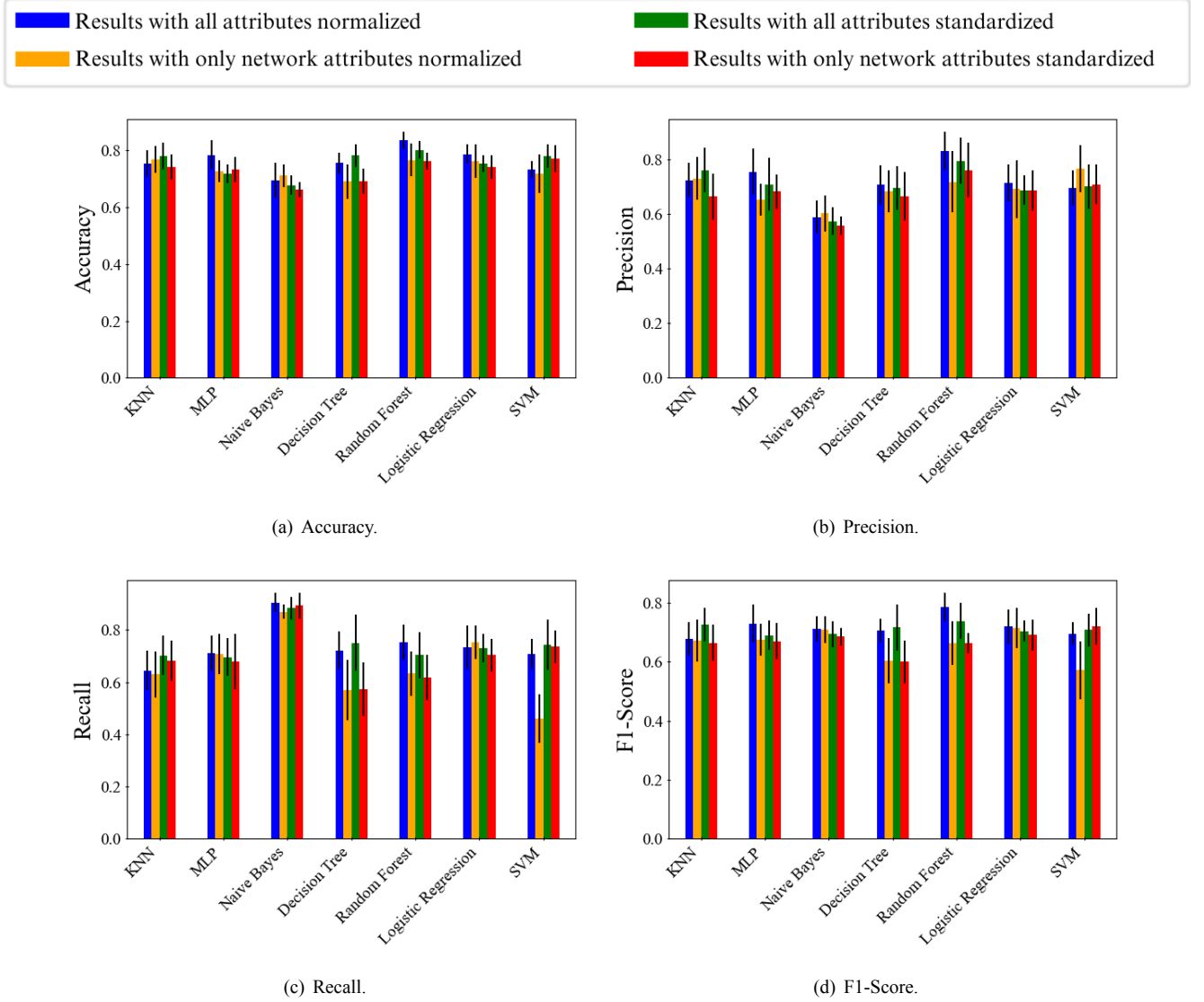


Figure 2. Comparison of the evaluated metrics to identify which machine learning model demonstrates the best performance on the considered dataset.

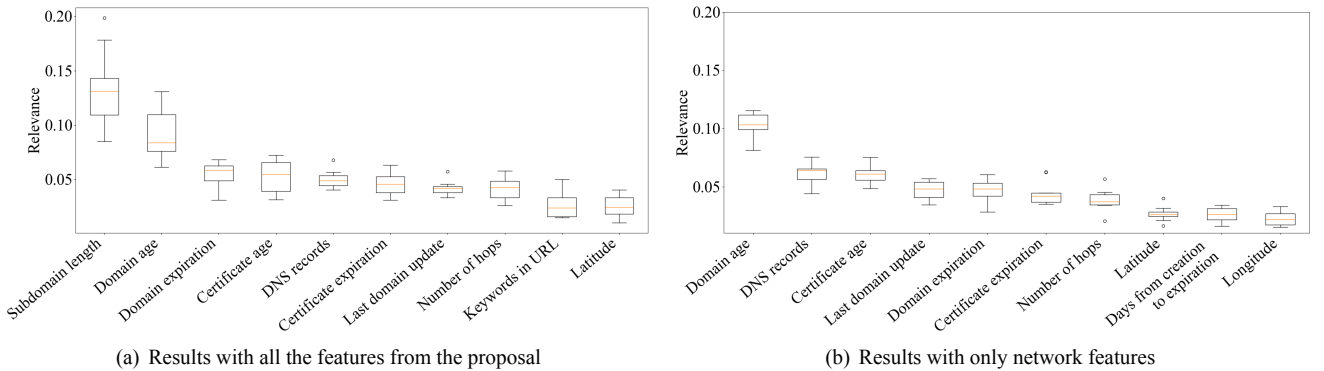


Figure 3. Comparison of the relevance distribution of a model trained with all the features reported in the proposal, using network features through box plots.

mains of websites to identify malicious domains. Saleem Raja *et al.* deployed machine learning models, reporting accuracy, precision, sensitivity, and F1-score results. Although the values presented in previous works are higher than those found in our work due to inherent differences, such as the size of the dataset and the choice of features, a similarity in the relationship between the models can be observed. Specif-

ically, the Random Forest achieved an accuracy of 0.98, precision of 0.99, sensitivity of 0.99, and F1-score of 1, while Naive Bayes obtained an accuracy of 0.79, precision of 0.84, sensitivity of 0.78, and F1-score of 0.77 [Saleem Raja *et al.*, 2021].

The proposed evaluation aims to observe the features most relevant for distinguishing between reliable and unreliable

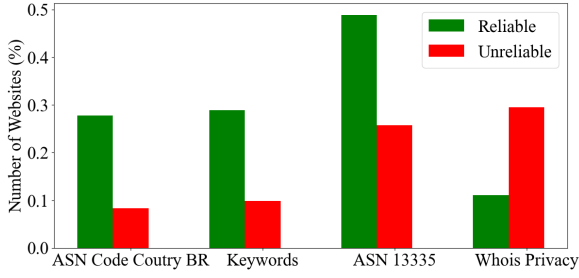


Figure 4. The top four categorical features, revealing the relationships between reliable and unreliable websites.

websites to understand the relationship between these features and the resulting classification. In this way, we used the data from the Random Forest with its normalized features. This is possible because the algorithm provides each feature's contribution percentage to the classification. Figure 3 compares the distribution of the ten most relevant features for the classification with all features and the classification with network features.

We conducted a dataset analysis focusing on the feature's relevance to comprehending the crucial features and their correlation with the websites' reliability. The most relevant features were characterized based on the contributions obtained from the Random Forest model to understand the relationship between the features and the classification. Figure 4 highlights the main categorical features of reliable and unreliable websites. The analysis also includes three additional highly relevant categorical features alongside those presented in Figure 3. Additionally, we created cumulative distribution plots to assess how the numerical features are distributed between the two sets, as shown in Figure 5. Figure 6 presents a heat map that evaluates geolocation-related features, such as Latitude and Longitude, by mapping the websites' locations across countries.

After analyzing the features, we have observed that the number of characters in a website's subdomain plays a crucial role in determining its reliability. The average number of characters in a subdomain for a reliable website is three (Figure 5(a)). We have also noticed that the age of the domain in days, as assessed by the domain creation days feature, is another essential factor. Reliable websites are generally older and require more effort to maintain their domain (Figure 5(b)). Furthermore, the number of days until the expiration of a domain is also a significant indicator of website reliability. Unreliable websites tend to use the domain for a shorter time than reliable ones, while reliable websites are more concerned with maintaining their domain for a more extended period. Figure 5(d) shows that reliable websites have longer domain update times. It happens because they contract domains to ensure a longer lifespan.

Notably, reliable websites usually have SSL/TLS certificates that last longer than unreliable ones. It is also evident that they hire certification services that provide certificates with more extended expiration periods. In contrast, unreliable websites may not have an SSL/TLS certificate. Additionally, to seem reliable, unreliable websites often use journalism-related keywords in their URL [Baly *et al.*, 2018; Rishikesh Mahajan, 2018].

The ASN 13335 refers to Cloudflare¹³, a globally recognized company that provides hosting and domain security services. The results, shown in Figure 4, indicate that nearly 50% of reliable websites use this service, while less than 30% of unreliable ones do. This behavior suggests that reliable websites have a greater tendency to choose more reliable services.

Furthermore, there is a notable trend among reliable websites to choose services in Brazil, as evidenced by analyzing the responsible organization's location based on ASN. This pattern indicates that unreliable websites tend to be outside the country and choose services abroad to avoid legal issues. Another validating result is the higher incidence of unreliable websites deciding to conceal domain registrant information. Figure 6 reveals that many unreliable websites tend to be located outside Brazil. While the results indicate the presence of many reliable websites outside Brazil, most are primarily based in the United States, mainly due to the availability of well-known hosting services and cloud providers. In contrast, unreliable websites show greater diversity, with locations in Canada and various European countries.

The number of TXT and CAA records (Figure 5(h)) evaluates the quantity of additional DNS records for domain security [Schwittmann *et al.*, 2019]. The obtained dataset shows that reliable websites tend to have a more significant number of records compared to unreliable ones.

Another critical feature is the number of hops (Figure 5(g)), which allows assessing the distance to the server when making a request, potentially resulting in higher network latency [Fisher, 2023]. The dataset shows unreliable websites tend to have more hops and, therefore, higher latency. The importance is also highlighted in the work of Couto *et al.*, which validates the occurrence of these features in reliable and unreliable websites. Couto *et al.* show that only 12.2% of unreliable websites are located in Brazil [Couto *et al.*, 2022]. Ahammad *et al.* use the LightGBM model to analyze the most relevant features, features such as subdomain and domain expiration also proved crucial in determining whether a website is malicious or not [Ahammad *et al.*, 2022].

Thus, it is observed that in both cases, domain-related features hold significant relevance in the model's classification, with particular emphasis on URL-related features and the number of keywords. However, when lexical features are removed from the results, relevance shifts, with the appearance of two new features: longitude and the time interval in days from the certificate's creation to its expiration.

The features engineered in this work rely on a previous work of Couto *et al.*, which characterizes high and low-credibility news websites without using machine learning models [Couto *et al.*, 2022]. However, our work uses a more up-to-date dataset than the one employed by Couto *et al.*. Our work also adds new features, such as the region code and continent of the website, the continent code of the autonomous system's location, whether the site uses HTTPS, and the encryption cipher. The following experiment compares the original features with the complete set of features proposed in our work, as shown in Table 4. The comparison

¹³Available at <https://www.cloudflare.com/>.

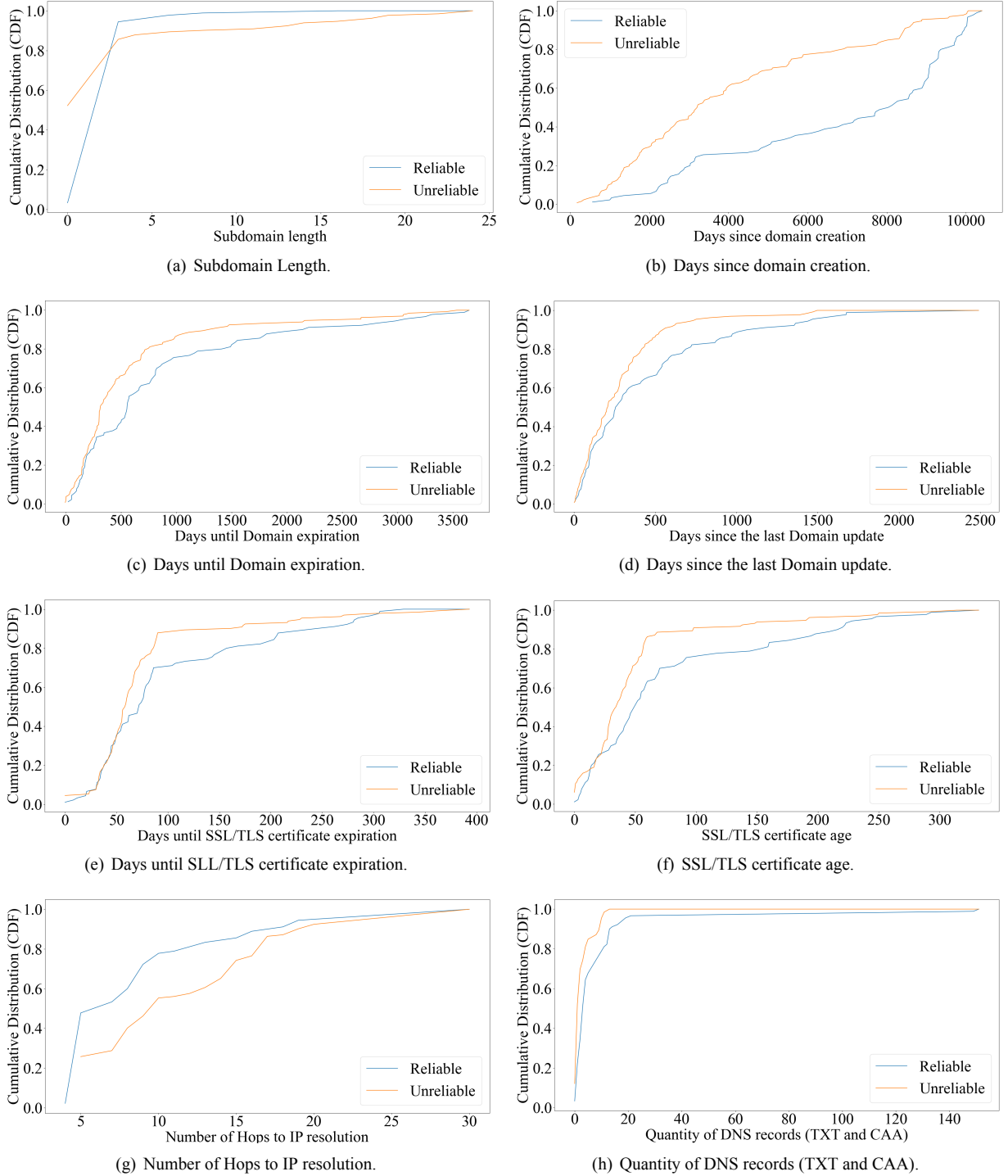


Figure 5. Analysis of the four most relevant features using information gain. Data comparing reliable and unreliable websites through cumulative distribution.

uses a Random Forest model with normalized features.

Comparing the results and considering the confidence interval, it is inferred that using domain and certification features is viable for classifying the reliability of news websites. However, the geolocation attributes lack the statistical significance needed to differentiate trusted websites from untrusted ones as effectively as other features. Consequently, the decision trees were assigned low relevance during the analysis. The dataset evaluation shows a similar distribution of websites within and outside Brazil across trusted and untrusted categories. This pattern arises from the concentration

of hosting services in global technology hubs, primarily outside Brazil.

The additional proposed features only marginally improve the model's performance. Therefore, these features do not provide the model with significantly more context about the data of each class, and their inclusion does not justify using more features for classifying websites.

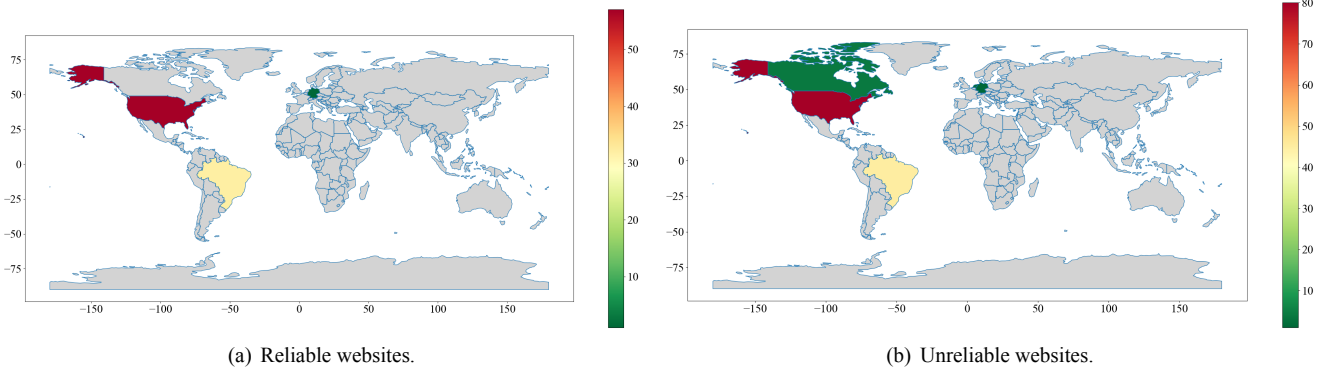


Figure 6. Comparison of the geolocation of reliable and unreliable websites through a heat map representing the number of websites by country.

Table 4. Comparison of the results obtained with the features identified in the foundational article by Couto *et al.* and with the features added in the proposal.

Model	Accuracy	Precision	Recall	F1-Score	AUC ROC
Proposal	0.84 ± 0.03	0.83 ± 0.07	0.75 ± 0.07	0.79 ± 0.05	0.91 ± 0.04
Couto <i>et al.</i>	0.81 ± 0.04	0.79 ± 0.07	0.72 ± 0.04	0.75 ± 0.05	0.9 ± 0.04

6 Conclusion

The spread of fake news and disinformation in Brazil profoundly impacts the political and social spheres, compromising the integrity of information.

This paper proposed an automatic approach with low computational resource commitment for classifying the reliability of websites disseminating information in the form of news. The proposal evaluated the reliability of Brazilian news websites through machine learning algorithms, considering features related to domain names, SSL/TLS certificate attributes, and geolocations. After normalizing all features, the Random Forest model performed better, achieving an average accuracy of 0.84.

Analysis of the results revealed that lexical features of the domain, such as subdomain size, the presence of keywords, and TLS certificate lifetime, were crucial in classification.

A limitation of this study is the dataset size. Given the confidence interval applied, the results may not generalize to larger datasets. Future work aims explore methods to expand the dataset, increasing the number of reliable and unreliable websites for improved robustness and representativeness. Future work also aims to explore hyperparameter optimization, data balancing, and variable selection to enhance machine learning models.

Understanding urban behavior is a complex task. In this work, we explored whether two different Location-Based Social Networks (LBSNs), Google Places and Foursquare, could provide comparable insights when modeling users' interests in geographic areas. We also examined how the modeling of LBSN data influences the definition and understanding of urban areas. Through an analysis of the characteristics of two datasets (resulting from Google Places and Foursquare LBSN data) and information collected from Curitiba, London, and several U.S. cities and counties, we found that the resulting graphs — referred to as Interest Networks (iNETs) — effectively capture the dynamics of users' behavior. These iNETs exhibit significant similarities, particularly in connections involving the most central urban areas at the

$h6$ granularity level, while greater differences emerge when smaller spatial units are employed for analysis (e.g., $h8$ and $h9$ granularity levels).

Additionally, we investigated whether LBSN users' interest in urban areas could be understood through geographic distance, political polarization, and socioeconomic characteristics from the areas they visit. Our findings indicate that, for the analyzed data, factors such as average income, racial composition, and political polarization of the areas (e.g., neighborhoods in Curitiba) do not sufficiently explain users' preferences. However, we observed that geographic distance plays a limiting role in interactions, as users tend to visit nearby areas.

Another aspect analyzed was the potential to minimize the differences observed between the iNETs derived from smaller areas (e.g., $h8$, $h9$). We proposed a method for defining (s) within a city, emphasizing the capture of densely connected areas. The iNETs formed from these zones, referred to as s, exhibited greater similarity across the two LBSNs analyzed.

It is important to highlight that the use of LBSNs must be accompanied by a critical understanding of the limitations and implications of their applicability. For instance, data from Google Places and Foursquare may not fully capture the interest of the entire population, as users of these platforms tend to be younger individuals with access to mobile internet. Nonetheless, these platforms provide valuable insights into the behavior of this demographic, which may or may not reflect broader societal patterns. Additionally, it is important to note that the data used in this study were collected a decade ago, meaning that the results might differ if more recent datasets were analyzed. However, the methodology remains applicable, and since the data sources used are publicly accessible, they offer a solid foundation for further exploration in urban computing research.

In future research, it would be valuable to analyze more recent datasets to determine whether different LBSNs continue to yield similar insights and to explore how factors influenc-

ing iNETs may change over time. This exploration could include examining additional variables such as cultural influences, types of venues, and the content of user reviews, including sentiment analysis and topic modeling. These factors would enrich the analysis and help address questions like whether the places frequented by users in urban environments share cultural similarities, whether individuals are inclined to visit the same types of venues across different areas, and whether certain regions are characterized by specific venue categories. Furthermore, it would be interesting to investigate whether the most interconnected areas within an iNET reflect similar sentiments among users. This comprehensive approach would enhance our understanding of urban behavior and the dynamics of city life concerning venues people commonly visit.

With a better understanding of the factors influencing the interests of the populations in each city, these analyses could be integrated into recommendation systems that highlight the unique characteristics of each neighborhood. Additionally, this type of investigation has the potential to inform public bodies about the most interconnected areas, enabling the development of public policies aimed at combating epidemics or promoting social integration between previously disconnected regions. Furthermore, a deeper exploration of the formation of , combined with validation from residents and experts in various cities, could enhance public policy systems for more effective management of available resources.

Declaration

Acknowledgements

We would like to express our sincere gratitude to the anonymous reviewers and the editorial team for their insightful comments and valuable suggestions, which have significantly improved the quality of this manuscript.

Funding

This work was funded by CNPq, CAPES, FAPERJ, RNP (Programa de bolsa de Incentivo à Pesquisa), Niterói City Hall/FEC/UFF (Edital PDPA 2020) and INCT ICONIOT.

Authors' Contributions

MPCM contributed to the conception, development, deployment, and writing of this study. IMM and DMFM contributed to the analysis of the results and writing. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets generated and analyzed during this study are available upon request by contacting the authors via email.

References

- Ahammad, S. H., Kale, S. D., Upadhye, G. D., Pande, S. D., Babu, E. V., Dhumane, A. V., and Bahadur, M. D. K. J. (2022). Phishing url detection using machine learning methods. *Advances in Engineering Software*, 173:103288. DOI: <https://doi.org/10.1016/j.advengsoft.2022.103288>.
- Al-Shehari, T. and Alsowail, R. A. (2021). An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques. *Entropy*, 23(10). DOI: 10.3390/e23101258.
- Alkawaz, M. H., Steven, S. J., Hajamydeen, A. I., and Ramli, R. (2021). A comprehensive survey on identification and analysis of phishing website based on machine learning methods. In *2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pages 82–87. DOI: 10.1109/ISCAIE51753.2021.9431794.
- Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., and Nakov, P. (2018). Predicting factuality of reporting and bias of news media sources. DOI: 10.48550/arXiv.1810.01765.
- BRASIL (2022). Tribunal superior eleitoral. programa permanente de enfrentamento à desinformação no âmbito da justiça eleitoral: plano estratégico: eleições 2022. *Biblioteca Digital da Justiça Eleitoral*. Available at: <https://bibliotecadigital.tse.jus.br/xmlui/handle/bdtse/9965>.
- Cordeiro, A., Sampaio, J., and Ruback, L. (2020). Fake-spread: Um framework para análise de propagação de fake news na web. In *Anais do XI Workshop sobre Aspectos da Interação Humano-Computador para a Web Social*, pages 9–16, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/wai-hews.2020.12342.
- Couto, J., Reis, J., Ítalo Cunha, Araújo, L., and Benvenuto, F. (2022). Caracterizando websites de baixa credibilidade no Brasil. In *Anais do XL Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 503–516, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbrc.2022.222361.
- de Oliveira, N. R., Pisa, P. S., Lopez, M. A., de Medeiros, D. S. V., and Mattos, D. M. F. (2021). Identifying fake news on social networks based on natural language processing: Trends and challenges. *Information*, 12(1). DOI: 10.3390/info12010038.
- Fisher, T. (2023). What are hops & hop counts?: What is a hop and why is it an important piece of information? Available at: <https://www.lifewire.com/what-are-hops-hop-counts-2625905>.
- Hua, J., Cui, X., Li, X., Tang, K., and Zhu, P. (2023). Multimodal fake news detection through data augmentation-based contrastive learning. *Applied Soft Computing*, 136:110125. DOI: 10.1016/j.asoc.2023.110125.
- Júnior, M., Melo, P., da Silva, A. P. C., Benvenuto, F., and Almeida, J. (2021). Towards understanding the use of telegram by political groups in brazil. In *Proceedings of the Brazilian Symposium on Multimedia and the Web, WebMedia '21*, page 237–244, New York, NY, USA. Association for Computing Machinery. DOI:

- 10.1145/3470482.3479640.
- Medeiros, D. S. V., Cunha Neto, H. N., Lopez, M. A., S. Magalhães, L. C., Fernandes, N. C., Vieira, A. B., Silva, E. F., and F. Mattos, D. M. (2020). A survey on data analysis on large-scale wireless networks: online stream processing, trends, and challenges. *Journal of Internet Services and Applications*, 11(1):6. DOI: 10.1186/s13174-020-00127-2.
- Palaniappan, G., S. S., Rajendran, B., Sanjay, Goyal, S., and B S, B. (2020). Malicious domain detection using machine learning on domain name features, host-based features and web-based features. *Procedia Computer Science*, 171:654–661. DOI: 10.1016/j.procs.2020.04.071.
- Posetti, J. and Matthews, A. (2018). A short guide to the history of ‘fake news’ and disinformation. *International Center for Journalists*, 7(2018). Available at: https://www.icfj.org/sites/default/files/2018-07/A%20Short%20Guide%20to%20History%20of%20Fake%20News%20and%20Disinformation_ICFJ%20Final.pdf.
- Ramos, M. d. M., Machado, R. d. O., and Cerqueira-Santos, E. (2022). “it’s true! i saw it on Whatsapp”: Social media, Covid-19, and political-ideological orientation in brazil. *Trends in Psychology*, 30(3). DOI: 10.1007/s43076-021-00129-4.
- Reis, J. C. S., Correia, A., Murai, F., Veloso, A., and Benvenuto, F. (2019). Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81. DOI: 10.1109/MIS.2019.2899143.
- Rishikesh Mahajan, I. S. (2018). Phishing website detection using machine learning algorithms. *International Journal of Computer Applications*, 181(23):45–47. DOI: 10.5120/ijca2018918026.
- Saleem Raja, A., Vinodini, R., and Kavitha, A. (2021). Lexical features based malicious url detection using machine learning techniques. *Materials Today: Proceedings*, 47:163–166. DOI: 10.1016/j.matpr.2021.04.041.
- Schwittmann, L., Wander, M., and Weis, T. (2019). Domain impersonation is feasible: A study of ca domain validation vulnerabilities. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 544–559. DOI: 10.1109/EuroSP.2019.00046.
- Sen, P. C., Hajra, M., and Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. In *Emerging Technology in Modelling and Graphics*, pages 99–111, Singapore. Springer Singapore. DOI: 10.1007/978-981-13-7403-6_11.
- Wardle, C. and Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking. Available at: <https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html>.
- Xuan, C. D., Nguyen, H. D., and Nikolaevich, T. V. (2020). Malicious url detection based on machine learning. *International Journal of Advanced Computer Science and Applications*, 11(1). DOI: 10.14569/IJACSA.2020.0110119.