Inferência Automática de Credibilidade de Sítios Web: Características de Domínio e Geolocalização para o Combate às Notícias Falsas

Marcos Paulo Cezar de Mendonça[1], **Igor Monteiro Moraes**[1] **e Diogo Menezes Ferrazani Mattos**[1]

[1] LabGen/MídiaCom – TET/IC/PPGEET/UFF
Universidade Federal Fluminense (UFF)
Niterói, RJ – Brasil

Resumo. *A avaliação da credibilidade de sítios web que propagam notícias é uma atividade crítica no combate à desinformação. Sítios web de baixa confiabilidade são por vezes apontados como a origem das notícias falsas propagadas e amplificadas em redes sociais. Este artigo propõe uma avaliação automática da credibilidade dos sítios web, sem a necessidade de varredura de todo o conteúdo do sítio. Diferente de trabalhos anteriores que focam nas redes sociais, este artigo utiliza características publicamente disponíveis dos sítios web, como as características do domínio, geolocalização e do certificado TLS, para identificar sítios web confiáveis e não confiáveis, usando técnicas de aprendizado de máquina supervisionado. O artigo propõe um modelo de aprendizado supervisionado e consolida um conjunto de dados de sítios confiáveis e não confiáveis. O modelo foi treinado e avaliado com dados disjuntos e foi possível identificar de forma eficaz, com precisão maior que 75%, sítios web confiáveis e não confiáveis, contribuindo para combate à disseminação de notícias falsas e de desinformação.*

Abstract. *Evaluating the credibility of websites that propagate news is a critical activity in combating disinformation. Websites of low reliability are sometimes pointed out as the origin of fake news propagated and amplified on social networks. This article proposes an automatic evaluation of the credibility of websites, without the need to scan all the site's content. Unlike previous works focusing on social networks, this article uses publicly available features of websites, such as domain characteristics, geolocation, and TLS certificate, to identify reliable and unreliable websites, using supervised machine learning techniques. The article proposes a supervised learning model and consolidates a dataset of reliable and unreliable sites. The model was trained and evaluated with disjoint data and it was possible to effectively identify, with an accuracy greater than 75%, reliable and unreliable websites, contributing to the fight against the spread of fake news and disinformation.*

1. Introdução

A disseminação de notícias falsas é um problema de pesquisa significativo que segue um padrão com características distintas de notícias confiáveis. As notícias falsas incluem manchetes persuasivas e textos projetados para atrair a atenção do

leitor[Posetti e Matthews, 2018]. De maneira informal, notícias falsas (*fake news*) e desinformação são comumente utilizados como sinônimos. Contudo, esses conceitos são formalmente diferentes. Wardle e Derakhshan definem *Fake News* como uma vasta gama de desinformação, *misinformation* e *malinformation*, na qual a desinformação são informações falsas que visam intencionalmente causar dano e confusão, enquanto *misinformation* são informações errôneas difundidas sem o objetivo de causar dano. A *malinformation* consiste na divulgação de fatos corretos, mas fora de contexto e com o propósito de causar dano. Ao comparar *fake news* e desinformação, nota-se que as *fake news* são projetadas para intencionalmente enganar o leitor, enquanto a desinformação é o resultado da disseminação de informações falsas, em diferentes meios, para enganar e manipular o público [Nemer, 2020, Wardle e Derakhshan, 2017, de Oliveira et al., 2021]. A grande quantidade de informações geradas diariamente torna a classificação de confiabilidade de um sítio web uma tarefa onerosa e inviável de ser executada manualmente, requerendo o auxílio de estratégias computacionais [Hua et al., 2023]. Muitos estudos atuais focam em abordagens para validar a aplicação de modelos de aprendizado de máquina na detecção de sítios com viés malicioso, como *phishing*, ou na identificação de disseminadores de notícias falsas [Alkawaz et al., 2021, Reis et al., 2019]. No entanto, a maioria das estratégias concentram-se na análise do conteúdo do sítio para a classificação da confiabilidade da informação.

Este trabalho propõe a classificação de sítios web confiáveis e não confiáveis por meio da aplicação de modelos de aprendizado de máquina supervisionados, como Naive Bayes, Máquina de Vetor de Suporte, Árvore de Decisão, Floresta Aleatória, Redes Neurais, Regressão Linear e K-Vizinhos Mais Próximos e através de atributos inerentes ao sítio web, como o domínio, certificado e geolocalização. Em comparação com trabalhos anteriores que se concentravam principalmente no conteúdo do site para detecção, a abordagem proposta foca na análise de características geográfica e do domínio dos sítios. Essas características são publicamente disponíveis e permitem a inferência da confiabilidade do site com baixo comprometimento de recursos computacionais. Os resultados obtidos demonstram a eficácia dessa abordagem, com a classificação de sítios confiáveis alcançando uma precisão maior que 75%.

O trabalho está organizado da seguinte forma. A Seção 2 elenca os trabalhos relacionados. A Seção 3 apresenta a proposta do trabalho. A Seção 4 descreve a metodologia de coleta de dados e de construção do conjunto de dados. A Seção 5 avalia a proposta e evidencia os principais resultados encontrados, comparando com a literatura. A Seção 6 conclui o trabalho.

2. Trabalhos Relacionados

O interesse crescente em combater os impactos da desinformação tem motivado pesquisadores, principalmente na área da computação, a propor alternativas eficazes contra a disseminação de desinformação [Santos et al., 2019]. Um dos enfoques para compreender a propagação de desinformação e notícias falsas é o *framework FakeSpread*, desenvolvido por Cordeiro *et al.*, que utiliza a teoria de grafos para analisar a disseminação a partir de uma fonte [Cordeiro et al., 2020]. O trabalho se baseia nos sítios web relatados na Comissão Parlamentar Mista de Inquérito (CPMI) das *Fake News*[1] do congresso

[1]Disponível em https://legis.senado.leg.br/comissoes/comissao?codcol=2292.

brasileiro para encontrar outros sítios que os utilizam como fonte de desinformação e, assim, construir uma relação entre os sítios e as fontes. Embora o enfoque do trabalho não seja a classificação da confiabilidade de um sítio web, mas em como a propagação é feita, o trabalho contribui com a ideia de que um sítio web não confiável utiliza fontes não confiáveis.

Outra abordagem, conduzida por Couto *et al.*, categoriza sítios web em baixa e alta credibilidade a partir de dados coletados no X (antigo Twitter), identificando usuários que postaram notícias como raiz [Couto et al., 2022]. Os autores coletam todas as notícias postadas anteriormente por esses usuários. A partir dessas fontes, atributos de certificação, registro e localização são obtidos. O trabalho não emprega modelos de aprendizado de máquina na automatização da classificação, mas permite inferir relações importantes entre sítios web de baixa e alta credibilidade por meio da comparação de seus atributos. Por sua vez, Baly *et al.* também utiliza atributos extraídos da URL dos sítios web para estudar notícias falsas e avaliar a classificação do viés político e a confiabilidade de um veículo de informação. Empregando aprendizado de máquina, foram extraídas amostras dos artigos publicados na Wikipedia e no X, além de informações do seu tráfego e da estrutura da URL [Baly et al., 2018]. Esses estudos reforçam a utilidade dos atributos retirados da estrutura da URL na validação da confiabilidade de uma informação.

No trabalho de Reis *et al.*, modelos de aprendizado de máquina são empregados para detectar notícias falsas a partir de artigos de notícias relacionados às eleições dos EUA de 2016, considerando atributos como viés político e localização do domínio [Reis et al., 2019]. Os dados obtidos são classificados de acordo com as principais características de sítios de baixa credibilidade presentes na literatura e propõe-se um novo conjunto de atributos, como o viés político, credibilidade e localização do domínio [Reis et al., 2019]. Saleem Raja *et al.* e Ahammad *et al.* focam a classificação de sítios web de *phishing*, utilizando atributos léxicos e de registro Whois, propondo a inclusão de atributos como certificado SSL/TLS e geolocalização [Saleem Raja et al., 2021, Ahammad et al., 2022]. Do Xuan *et al.* e Palaniappana *et al.* também exploram atributos léxicos da URL para classificar sítios maliciosos, empregando modelos de aprendizado de máquina como Floresta Aleatória e Regressão Logística, respectivamente [Do Xuan et al., 2020, Palaniappan et al., 2020].

Os trabalhos anteriores abordaram a classificação de sítios web maliciosos e a detecção de notícias falsas, utilizando uma variedade de abordagens, como análise de grafos, atributos léxicos da URL e modelos de aprendizado de máquina. No entanto, muitos desses estudos se concentraram na classificação de conteúdo com base em dados de redes sociais ou em análises detalhadas do texto das notícias. Em contraste, a presente proposta diferencia-se ao focar na avaliação automática da credibilidade do sítio web em si e na identificação da origem das notícias falsas diretamente a partir das características do domínio, geolocalização e certificado SSL/TLS. Isso elimina a necessidade de varredura de todo o conteúdo do sítio, proporcionando uma abordagem mais eficiente e direta para mitigar a disseminação da desinformação *online*.

3. Proposta de Classificação de Confiabilidade de Sítios Web

A proposta deste trabalho é aplicar os algoritmos de aprendizado de máquina supervisionados na classificação de sítios web de notícias confiáveis e não confiáveis através de atributos de redes. Os algoritmos de aprendizado de máquina supervisionado lidam com dados que são previamente categorizados juntamente com seus atributos e os utiliza para treinamento e aprendizado de padrões. Padrões que são posteriormente empregados para a classificação de categorias ou previsão de valores [Sen et al., 2020].

Para a classificação com algoritmos de aprendizado de máquina, é necessário criar um conjunto de dados que será utilizado para treinar os algoritmos por meio de exemplos. Esse conjunto de dados deve conter atributos que descrevam a categoria de cada registro. Considerando a escolha dos atributos que representem as características dos sítios web com o intuito de classificar entre confiáveis e não confiáveis, muitos trabalhos utilizam atributos extraídos das análise do conteúdo do sítio web ou da relação com as redes sociais. Neste trabalho, seguiu-se a metodologia de Couto *et al.* que visa caracterizar a relação da confiabilidade de sítios web de notícias brasileiros e atributos divididos em três categorias: domínio, certificado e geolocalização [Couto et al., 2022]. Atributos que possuem um menor custo computacional e, portanto, exigem menos recursos e tempo para serem obtidos.

Os atributos de domínio incluem aqueles relacionados ao registro DNS (*Domain Name System*) que mapeia uma URL (*Uniform Resource Locator*) a um endereço IP (*Internet Protocol*). Durante o processo de registro de um domínio, diversas informações são necessários para a sua configuração. Para registrar um domínio, deve-se escolher um serviço de hospedagem e seus servidores, gerando, assim, informações importantes que caracterizam aquele sítio. Esses atributos contém as informações sobre o subdomínio; os dias de criação, expiração e atualização do domínio; a identidade de quem fez o registro do domínio. Outro atributo importante obtido nessa categoria é o número de sistema autônomo (*Autonomus System Number* - ASN) que referencia um sistema autônomo dirigido por uma organização e que hospeda o sítio web. Esses sistemas autônomos compõem a camada de rede na Internet e são responsáveis pelo roteamento entre outros sistemas autônomos e o gerenciamento de seus endereços IPs. Além disso, são utilizados dados relativos à estrutura da URL, como o tamanho do nome do subdomínio, o domínio de nível mais alto (*Top-Level Domain* - TLD) e a se existem palavras-chave relacionadas ao escopo de notícias.

Os atributos de certificado são aqueles relacionados ao certificados SSL/TLS que um sítio web pode possuir. Sítios web que fazem uso do protocolo HTTPS (*Hypertext Transfer Protocol Secure*) acrescentam uma camada de segurança para a transmissão segura dos dados entre o cliente e o servidor através da criptografia da informação. Os certificados SSL/TLS podem ser obtido para qualquer domínio e contêm informações importantes sobre quem os emitiu, quando é sua data de expiração, quando foi criado e a entidade responsável por gerá-los. Os atributos de geolocalização estão relacionados à localização física onde estão hospedados os servidores do sítio web. Os atributos extraídos descrevem em que país e em qual região se encontram os servidores associados aos sítios, além da informação de geolocalização associada ao ASN.

Além dos atributos elencados por Couto *et al.*, outros atributos foram adicionados por este trabalho para dar aos modelos de aprendizado de máquina um maior contexto

sobre os sítios web. Os atributos adicionais são: um atributo que verifica se o sítio web, ao ser acessado, redireciona para a porta HTTPS, qual a cifra utilizada de encriptação utilizada pelo certificado SSL/TLS, a região do país onde se localiza o endereço IP e o ASN. Ademais, este trabalho também procura estudar a classificação dos sítios web confiáveis e não confiáveis com atributos de redes que independam de sua estrutura léxica do sítio web.

4. Metodologia de Avaliação

Para avaliar a proposta deste trabalho, aplicou-se a metodologia descrita na Figura 1, na qual, divide-se em três etapas. Na primeira etapa, é feita a construção do conjunto de dados (*dataset*), buscando as URLs dos sítios confiáveis e não confiáveis. Após isso, é realizada a obtenção dos atributos para cada sítio. A segunda etapa lida com o pré-processamento do conjunto de dados para preparar os dados para os modelos de aprendizado de máquina. Na terceira e última etapa, realiza-se a escolha dos hiperparâmetros para cada modelo e a classificação.

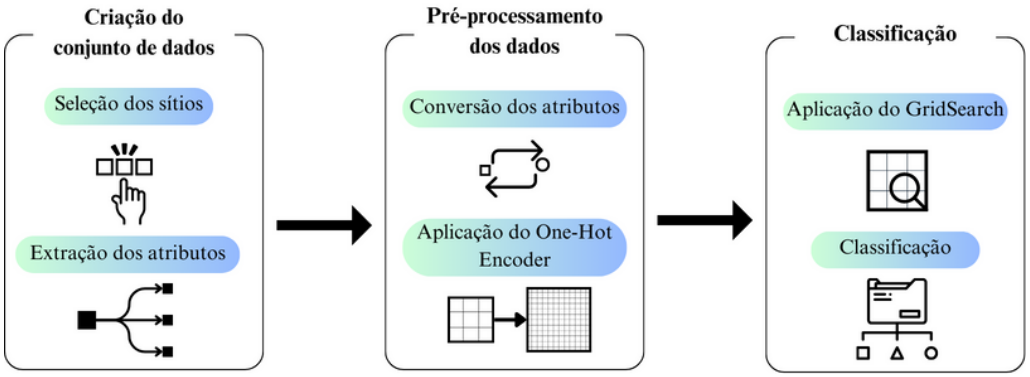


Figura 1. Fluxo das etapas da metodologia proposta para a classificação de sítios web confiáveis e não confiáveis.

4.1. Criação do Conjunto de Dados

A criação do conjunto de dados para a classificação é realizada em duas etapas. A primeira consiste na seleção de sítios web de notícias confiáveis e não confiáveis. A partir desta seleção, são extraídos os atributos de rede que representam aquele sítios. Os sítios web confiáveis foram extraídos da listagem dos sítios associados à Associação Nacional de Jornais (ANJ)[2], uma associação brasileira que defende os interesses dos jornais e contribui para o desenvolvimento dos jornais através da troca de experiências, da disseminação de inovações e da cooperação entre empresas e entidades semelhantes. Ao todo foram obtidos 95 sítios web, porém cinco destes estavam fora dor ar no momento da pesquisa. Os sítios web não confiáveis foram obtidos a partir da metodologia definida por Cordeiro *et al.*. No trabalho, os autores estudam a relação de como sítios web não confiáveis referenciam em outros sítios não confiáveis [Cordeiro et al., 2020]. Para isso, foram usados os sítios relatados na CPMI da *Fake News* para buscar mais sítios que os citam. Para tanto, foi usada a `API Custom Search JSON`[3] do Google que fornece um

[2]Disponível em https://www.anj.org.br/.

[3]Disponível em https://developers.google.com/custom-search/v1/overview?hl=pt-br.

buscador personalizado capaz de realizar pesquisas através de requisições HTTP. Assim, para cada sítio relatado na CPMI, buscou-se os 100 primeiros, caso houvesse, sítios web de notícias que citavam como fonte um dos sítios apontados pela CPMI da *Fake News*. Ao todo foram obtidos 240 sítios web e, então, foi realizada uma filtragem para remover domínios que pertenciam ao escopo de portais de notícias como redes sociais, referências duplicadas ou referências a arquivos, como PDFs. Ademais, foi feita a remoção dos sítios que não estavam disponíveis no momento de realização da pesquisa, totalizando 132 sítios web rotulados como não confiáveis.

A partir dos sítios web selecionados, foi possível extrair os atributos das categorias descritos na Seção 3. Para obter os atributos de domínio, tais como a existência de hífen no domínio ou o tamanho do subdomínio, primeiramente foi realizada uma análise sobre a estrutura da URL do sítios. Os atributos relacionados aos dados do dono do domínio, disponibilizados por serviços `WHOIS`, foram extraídos com a biblioteca Python Whois[4]. Os atributos relacionados ao registro DNS, como o número de registros dos tipos `TXT` e `CAA` foram obtidos através de requisições HTTP à API Ninja [5], através de uma chave de acesso gratuita. Obteve-se o número de saltos (*hops*) utilizando o comando `tracert`. A API IPWHOIS.IO[6], foi utilizada para obter as informações relativas aos dados do ASN e as geolocalizações dos servidores. Para a extração de atributos de certificado, empregaram-se bibliotecas da própria linguagem Python, como a SSL[7] e Socket[8].

4.2. Pré-Processamento dos Dados

A primeira parte deste pré-processamento é transformar todos os dados antes salvos como valores *booleanos* na linguagem Python (*True* ou *False*) em valores numéricos (1 ou 0) com auxílio da biblioteca Pandas[9]. Na segunda parte do pré-processamento, o método *One-Hot Encode* foi empregado para lidar com os dados categóricos. Esse método, como visto na Figura 2, consiste em transformar cada valor distinto de um atributo categórico em uma coluna e atribui o valor 1, caso o registro apresente aquele valor [Al-Shehari e Alsowail, 2021].

	Atributo 1	Atributo 2
Dado 1	1	A
Dado 2	2	B
Dado 3	3	C

Aplicar *One-Hot Encoding* →

	Atributo 1	Atributo 2_A	Atributo 2_B	Atributo 2_C
Dado 1	1	1	0	0
Dado 2	2	0	1	0
Dado 3	3	0	0	1

Figura 2. Exemplo de aplicação do *One-Hot Encoding* sobre um conjunto de dados. O método é responsável por transformar valores de um atributo categórico em novas colunas.

Para implementar este método, identificaram-se todos os atributos categóricos, que no escopo deste trabalho, são todas as colunas contendo dados não-numéricos, além

[4]Disponível em https://github.com/richardpenman/whois.
[5]Disponível em https://api-ninjas.com/api/dnslookup.
[6]Disponível em https://ipwhois.io/.
[7]Disponível em https://docs.python.org/3/library/ssl.html.
[8]Disponível em https://docs.python.org/3/library/socket.html.
[9]Disponível em https://pandas.pydata.org/.

da coluna de ASN. Embora o ASN seja um número, optou-se em realizar o *One-Hot Encode* para evitar qualquer tentativa dos modelos de atribuírem ordem ou magnitude a esses valores. Com o auxílio do módulo do OneHotEncoder do Scikit-learn [10], realizou-se esse processamento.

4.3. Classificação dos Sítios Web

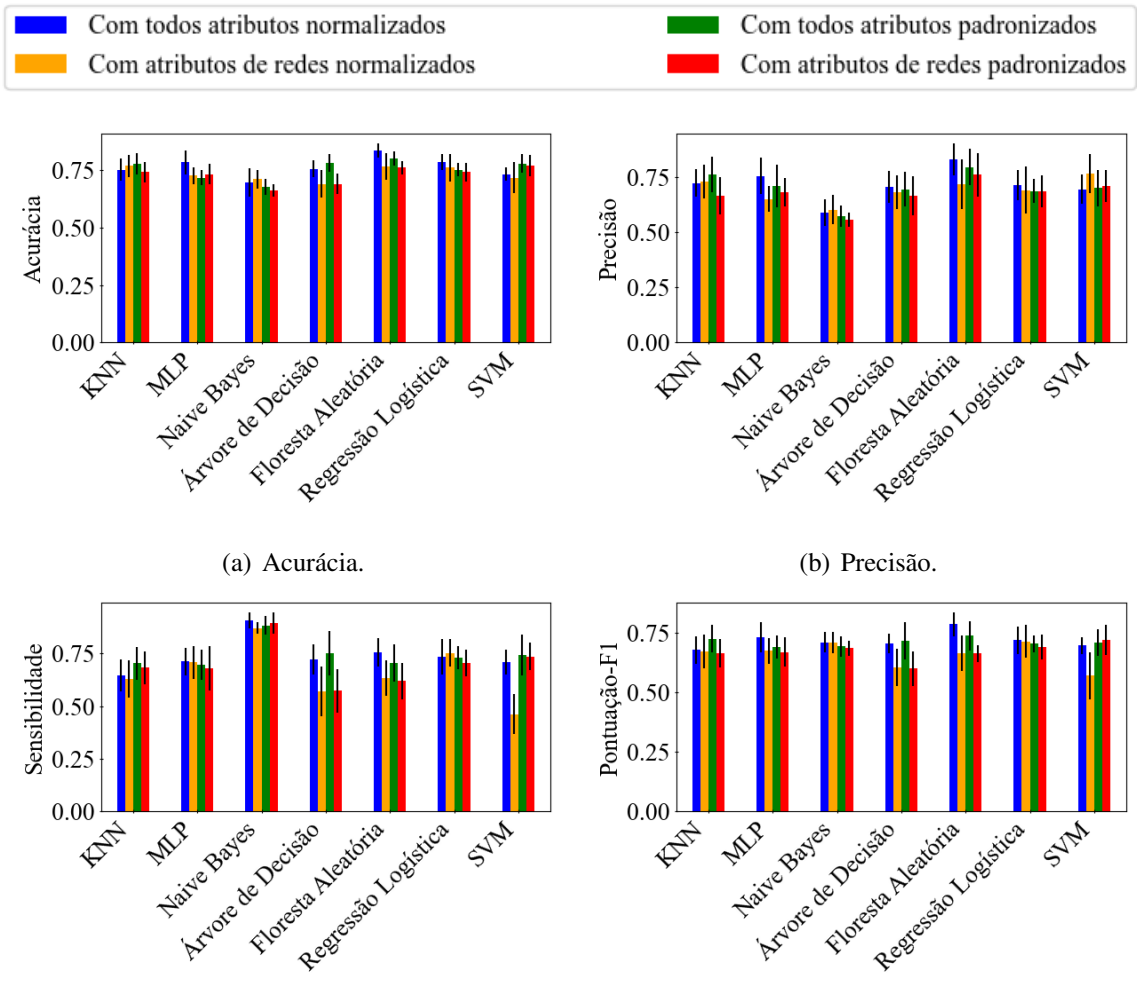
Com o conjunto de dados pré-processado, a etapa seguinte é o treinamento e teste para cada modelo de aprendizado de máquina para avaliar os resultados. Durante essa etapa, optou-se por realizar tanto a padronização quanto a normalização dos dados para cada modelo, resultando em dois conjuntos diferentes. Para efetuar a padronização e a normalização, bem como o treinamento e teste dos modelos, foi empregado o módulo *Pipeline* da biblioteca Scikit-learn [11]. Esse módulo permite definir uma sequência de operações a serem realizadas. O Pipeline retorna um objeto utilizado para treinar tanto o padronizador, ou o normalizador, quanto o modelo. Após a construção do *Pipeline*, foram realizadas 10 repetições para obtenção dos resultados. Em cada repetição, o conjunto de dados foi dividido em dois conjuntos distintos, um de treinamento, composto por 80% dos dados originais, e um de teste, com os 20% restantes. Durante a primeira repetição de cada modelo, implementou-se a busca em grade (*gridsearch*) sobre os dados de treinamento para obter os melhores hiperparâmetros do modelo de aprendizado. Após definir um conjunto de valores hiperparâmetros, aplicou-se o *gridsearch* sobre os dados. Foram realizadas mais 10 repetições, nas quais os dados de treinamento foram divididos em 90% para treinamento dentro do *gridsearch* e 10% para validação. Em cada repetição, o modelo com uma combinação possível de hiperparâmetros foi treinado e avaliado. Os hiperparâmetros que possuírem a melhor média de área abaixo da curva (*Area Under the Curve* - AUC) característica de operação do receptor (*Receiver Operating Characteristic* - ROC). Com os hiperparâmetros definidos, realizaram-se o treinamento e a predição em cada repetição. Na etapa de predição, os dados de teste, sem a classificação, foram passados para o modelo, que determinou a classificação de cada registro com base nos dados de treinamento. A partir dos resultados dessa predição e das classificações reais do conjunto de teste, construiu-se a matriz de confusão e extraíram-se as métricas avaliadas neste trabalho.

5. Avaliação dos Resultados

A avaliação do desempenho de cada modelo de aprendizado de máquina para a detecção de sítios web confiáveis e não confiáveis considera os atributos de domínio, certificado e geolocalização extraídos da URL de cada sítio. Com isso, é possível avaliar qual modelo melhor se adapta à proposta. Além disso, para cara modelo, as métricas foram calculadas com todos os atributos definidos na proposta e, também, considerando somente os atributos de redes de cada sítio, removendo os atributos léxicos da URL do conjunto de dados. Também foram realizados estudos para verificar a relevância dos atributos para classificação e a comparação utilizando somente os apresentados no trabalho de Couto *et al.* [Couto et al., 2022]. Para avaliar os modelos que melhor se adequaram ao conjunto de dados, a média da acurácia, precisão, sensibilidade e a pontuação-F1 são comparadas

[10]Disponível em https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html.

[11]Disponível em https://scikit-learn.org/stable/modules/compose.html.



(a) Acurácia.

(b) Precisão.

(c) Sensibilidade.

(d) Pontuação-F1.

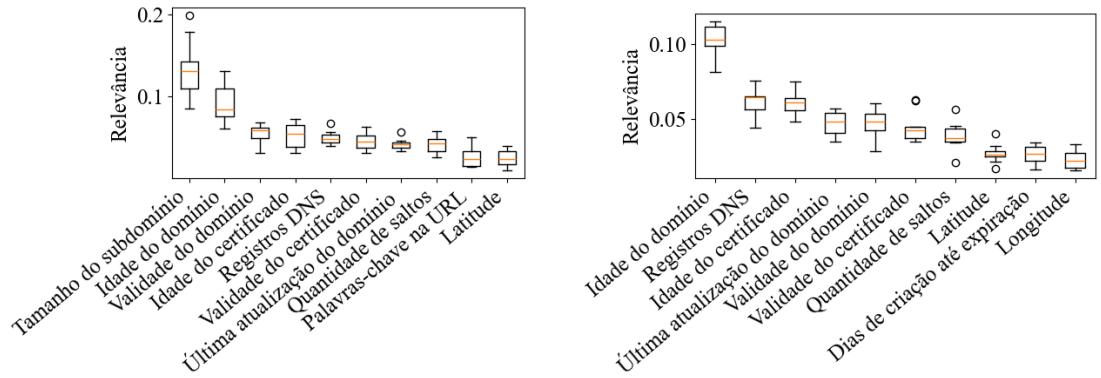
Figura 3. Comparação das métricas avaliadas para identificar qual modelo de aprendizado de máquina apresenta melhor desempenho no conjunto de dados considerado.

entre os modelos, considerando um intervalo de confiança de 95%. Os resultados são mostrados na Figura 3.

De acordo com a Figura 3(a) a acurácia de todos os modelos, com exceção do algoritmo Naive Bayes, está acima de 70% para os métodos que consideram todos os atributos, com destaque para a Floresta Aleatória, tanto com normalização quanto com padronização, 77% e 76% de média. Os modelos que foram treinados apenas com atributos de redes tiveram, em sua maioria, médias abaixo de 70%. O mesmo comportamento se repete ao analisar os resultados obtidos pela precisão, vistos na Figura 3(b), e a pontuação-F1, Figura 3(d). Esses resultados mostram que o modelo que apresenta melhor desempenho sobre conjunto de dados considerado foi a Floresta Aleatória com todos os atributos. Isso ocorre devido à sua características de construir diversas árvores de decisão de forma a dar mais relevâncias aos atributos que reduzem a perda logarítmica, sendo assim, possível de identificar atributos mais relevantes e dar maiores pesos para eles. Os resultados obtidos pelo Naive Bayes podem ser explicados pelo fato de o conjunto de

dados considerado ser esparso, apresentando mais de 200 atributos, com atributos que não convergem para a alta probabilidade de decisão, já que há atributos relacionados às classes que aparecem no conjunto de treinamento e não no de teste. Diferente dos resultados anteriores, o Naive Bayes foi o modelo que obteve maior sensibilidade, como pode ser visto na Figura 3(c). Logo, esse modelo conseguiu ter uma quantidade de acertos maior considerando somente a classe positiva. Esse resultado justifica-se pelo fato de a estrutura do conjunto de dados ter uma variedade maior de valores para os sítios não confiáveis do que para os confiáveis, fazendo com que fosse possível relacionar os dados entre o conjunto de treinamento e teste. Ao comparar os resultados obtidos com todos os atributos da proposta e dos atributos de redes, nota-se que os resultados que utilizaram todos os atributos obtiveram melhor desempenho de classificação. Logo, os atributos léxicos da URL possibilitam aos modelos terem um maior contexto sobre os dados do sítio web, e assim, ter uma capacidade de classificação melhor. Os resultados deste estudo corroboram trabalhos anteriores [Reis et al., 2019, Saleem Raja et al., 2021]. Reis *et al.* destacam que a Floresta Aleatória superou o Naive Bayes [Reis et al., 2019]. Saleem Raja *et al.* usaram atributos da URL e do domínio para identificar domínios maliciosos, alcançando resultados semelhantes [Saleem Raja et al., 2021].

A avaliação proposta visa observar os atributos que são mais relevantes para a distinção dos sítios web confiáveis e não confiáveis para entender a relação desses atributos com o conjunto de dados e a classificação resultante. Para isso, utilizaram-se os dados da Floresta Aleatória com seus atributos normalizados. Isso é possível, pois o próprio algoritmo, tem como resposta, a porcentagem de contribuição de cada atributo para a classificação. A Figura 4 mostra a comparação da distribuição dos dez atributos mais relevantes tanto para a classificação com todos os atributos quanto para a classificação com os atributos de rede.



(a) Resultado com todos os atributos da proposta

(b) Resultado com atributos de rede

Figura 4. Comparação da distribuição da relevância de um modelo treinado com todos os atributos relatados na proposta quando os atributos de rede a partir de diagramas de caixa.

Analisando os atributos, tem-se que a quantidade de caracteres no subdomínio de um sítio é relevante para a definição da confiabilidade. A média para um sítio web confiável é três caracteres. Os atributos relacionados aos dias de criação avaliam a idade do domínio em dias. Sítios web confiáveis costumam ser mais antigos e terem mais

esforços aplicados na manutenção daquele domínio. A quantidade de dias até a expiração faz alusão à ideia de que sítios web não confiáveis utilizam o domínio por um tempo menor que os confiáveis. Por sua vez, sítios confiáveis tendem a ter maior preocupação com a manutenção do domínio. Do mesmo modo, sítios web confiáveis tendem a ter um certificado SSL/TSL há mais tempo que os não confiáveis, além de contratarem serviços de certificação que proveem certificados com tempos até a expiração mais longos. Há ainda a possibilidade de não possuírem certificado SSL/TLS. Os atributo de latitude e longitude estão relacionados à localização geográfica do servidor. Sítios web de notícias não confiáveis tendem a ter localizações fora do Brasil para evitarem problemas jurídicos. A existência de palavras-chave relacionadas ao jornalismo é verificada na URL, pois sítios web não confiáveis tendem a utilizar essas palavras para aparentar confiabilidade [Baly et al., 2018, Mahajan e Siddavatam, 2018].

O número de registros `TXT` e `CAA` avalia a quantidade de registros DNS adicionais para a segurança do domínio [Schwittmann et al., 2019]. A partir do conjunto de dados obtidos, tem-se que em sítios web confiáveis, a quantidade de registros tendem a ser maior que os não confiáveis. Outro atributo crítico é a quantidade de saltos que permite avaliar a distância até o servidor na realização da requisição, o que pode ocasionar maior latência na rede[Fisher, 2023]. Nota-se que, a partir do conjunto de dados, que sítios web não confiáveis tendem a ter uma maior quantidade de saltos e, portanto, tendem a ter maior latência. Observa-se ainda essa importância também no trabalho de Couto *et al.* que valida a incidência desses atributos em sítios confiáveis e não confiáveis. Couto *et al.* mostram que apenas 12,2% dos sítios web não confiáveis se encontram no Brasil[Couto et al., 2022]. No estudo realizado por Ahammad *et al.*, no qual, os autores analisam através do modelo LightGBM os atributos mais relevantes, atributos como o subdomínio e expiração do domínio também mostraram ser importantes para determinar se um sítio web é malicioso ou não [Ahammad et al., 2022]. Sendo assim, nota-se que nos dois casos, os atributos relacionados ao domínio possuem grande relevância na classificação gerada pelo modelo, com destaque para os atributos relacionados à URL e à quantidade de palavras-chave. No entanto, nos resultados que removem os atributos léxicos, há uma mudança na ordem de relevância e a aparição de dois atributos, como a Longitude e o intervalo em dias desde a criação até a expiração do certificado.

Os atributos utilizados neste trabalho são baseados no estudo de Couto *et al.* que visa estudar a caracterização de sítios web de notícias de alta e baixa credibilidade sem a utilização de modelos de aprendizado de máquina [Couto et al., 2022]. Foi empregado um conjunto de dados mais atualizado que o empregado por Couto *et al.* São considerados novos atributos, como código da região e continente do sítio web, além do código do continente onde está localizado o sistema autônomo, a execução do sítio sobre HTTPS e a cifra de encriptação. Assim, realiza-se um estudo comparativo entre os atributos originais e o conjunto total de atributos proposto neste trabalho, presente na Tabela 1. A comparação é realizada utilizando a Floresta Aleatória com normalização dos atributos.

Tabela 1. Comparação dos resultados obtidos com atributos levantados pelo artigo base de Couto *et al.* e com os atributos adicionados na proposta.

Modelo	Acurácia	Precisão	Sensibilidade	Pontuação-F1	AUC ROC
Proposta	0.84±0.03	0.83±0.07	0.75±0.07	0.79±0.05	0.91±0.04
Couto *et al.*	0.81±0.04	0.79±0.07	0.72±0.04	0.75±0.05	0.9±0.04

A partir da comparação dos resultados, considerando a margem de erro, infere-se a viabilidade de utilizar os atributos de domínio e geolocalização para a classificação da confiabilidade de sítios de notícias, embora a utilização dos atributos adicionais propostos contribuam de maneira marginal para a melhora do desempenho do modelo. Logo, os atributos adicionais não permitem ao modelo ter um maior contexto sobre os dados de cada classe, não justificando o uso de mais atributos para a classificação dos sítios web.

6. Conclusão

A disseminação de notícias falsas no Brasil impacta profundamente as esferas políticas e sociais, comprometendo a integridade das informações. Esse trabalho propôs uma abordagem automática e com baixo comprometimento de recursos computação para a classificação da confiabilidade de sítios web que disseminam informações na forma de notícias. A proposta avaliou a confiabilidade de sítios web de notícias brasileiras através de algoritmos de aprendizado de máquina, considerando características relativas aos nomes de domínio, certificados e geolocalizações. O modelo de Floresta Aleatória, com todos os atributos normalizados, demonstrou melhor desempenho, alcançando uma acurácia média de 0,84. A análise dos resultados revelou que atributos léxicos do domínio, como tamanho do subdomínio e presença de palavras-chave, e características de geolocalização, como latitude do servidor registrado e tempo de vida do certificado TLS, foram cruciais na classificação.

Os trabalhos futuros visam explorar a otimização de hiperparâmetros, balanceamento de dados e seleção de variáveis para aprimorar os modelos de aprendizado de máquina.

Referências

[Ahammad et al., 2022] Ahammad, S. H., Kale, S. D., Upadhye, G. D., Pande, S. D., Babu, E. V., Dhumane, A. V. e Bahadur, M. D. K. J. (2022). Phishing url detection using machine learning methods. *Advances in Engineering Software*, 173:103288.

[Al-Shehari e Alsowail, 2021] Al-Shehari, T. e Alsowail, R. A. (2021). An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques. *Entropy*, 23(10):1258.

[Alkawaz et al., 2021] Alkawaz, M. H., Steven, S. J., Hajamydeen, A. I. e Ramli, R. (2021). A comprehensive survey on identification and analysis of phishing website based on machine learning methods. Em *2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, p. 82–87. IEEE.

[Baly et al., 2018] Baly, R., Karadzhov, G., Alexandrov, D., Glass, J. e Nakov, P. (2018). Predicting factuality of reporting and bias of news media sources. *arXiv preprint arXiv:1810.01765*.

[Cordeiro et al., 2020] Cordeiro, A., de Oliveira Sampaio, J. e Ruback, L. (2020). Fakespread: Um framework para análise de propagação de fake news na web. Em *Anais do XI Workshop sobre Aspectos da Interação Humano-Computador Para a Web Social*, p. 9–16. SBC.

[Couto et al., 2022] Couto, J. M., Reis, J. C., Cunha, Í., Araújo, L. e Benevenuto, F. (2022). Caracterizando websites de baixa credibilidade no Brasil. Em *Anais do XL Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, p. 503–516. SBC.

[de Oliveira et al., 2021] de Oliveira, N. R., Pisa, P. S., Lopez, M. A., de Medeiros, D. S. V. e Mattos, D. M. F. (2021). Identifying fake news on social networks based on natural language processing: Trends and challenges. *Information*, 12(1).

[Do Xuan et al., 2020] Do Xuan, C., Nguyen, H. D. e Tisenko, V. N. (2020). Malicious url detection based on machine learning. *International Journal of Advanced Computer Science and Applications*, 11(1).

[Fisher, 2023] Fisher, T. (2023). What are hops & hop counts?: What is a hop and why is it an important piece of information?

[Hua et al., 2023] Hua, J., Cui, X., Li, X., Tang, K. e Zhu, P. (2023). Multimodal fake news detection through data augmentation-based contrastive learning. *Applied Soft Computing*, 136:110125.

[Mahajan e Siddavatam, 2018] Mahajan, R. e Siddavatam, I. (2018). Phishing website detection using machine learning algorithms. *International Journal of Computer Applications*, 181(23):45–47.

[Nemer, 2020] Nemer, D. (2020). Desinformação no contexto da pandemia do coronavírus (covid-19). *AtoZ: novas práticas em informação e conhecimento*, 9(2):113–116.

[Palaniappan et al., 2020] Palaniappan, G., Sangeetha, S., Rajendran, B., Goyal, S., Bindhumadhava, B. et al. (2020). Malicious domain detection using machine learning on domain name features, host-based features and web-based features. *Procedia Computer Science*, 171:654–661.

[Posetti e Matthews, 2018] Posetti, J. e Matthews, A. (2018). A short guide to the history of 'fake news' and disinformation. *International Center for Journalists*, 7(2018).

[Reis et al., 2019] Reis, J. C., Correia, A., Murai, F., Veloso, A. e Benevenuto, F. (2019). Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81.

[Saleem Raja et al., 2021] Saleem Raja, A., Vinodini, R. e Kavitha, A. (2021). Lexical features based malicious url detection using machine learning techniques. *Materials Today: Proceedings*, 47:163–166. NCRABE.

[Santos et al., 2019] Santos, W. R., Xavier, M. R., da Cunha, D. C., Júnior, J. C., Adauto, D. A. e Ferraz, C. A. (2019). Trendsbot: Verificando a veracidade das mensagens do telegram utilizando data stream. Em *Anais Estendidos do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, p. 65–72. SBC.

[Schwittmann et al., 2019] Schwittmann, L., Wander, M. e Weis, T. (2019). Domain impersonation is feasible: A study of ca domain validation vulnerabilities. Em *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, p. 544–559.

[Sen et al., 2020] Sen, P. C., Hajra, M. e Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. Em *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018*, p. 99–111. Springer.

[Wardle e Derakhshan, 2017] Wardle, C. e Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg.