Caracterizando Websites de Baixa Credibilidade no Brasil

João M. M. Couto[1], Julio C. S. Reis[2], Ítalo Cunha[1], Leandro Araújo[1], Fabrício Benevenuto[1]

[1]Universidade Federal de Minas Gerais (UFMG) -– Brasil

[2]Universidade Federal de Viçosa (UFV) -– Brasil

{joaocouto,cunha,leandroaraujo,fabricio}@dcc.ufmg.br,jreis@ufv.br

Abstract. *Misinformation is a growing problem across the globe particularly, in Brazil. The ease of information spreading promoted by the adoption of digital platforms severely intensified the extend of this problem. However, in mosts cases, these platforms are utilized purely as means of dissiminating content actually produced by external websites designed specifically for the creation and spread of fake news. Therefore, in this effort, we present a extensive characterzation of high and low credibility websites. Our results reveal interesting characteristics of low credibility websites that may also be useful to distinguish them from other websites, thus furthering our understanding of misinformation in Brazil.*

Resumo. *Desinformação é um problema crescente no mundo e, em particular, no Brasil. A facilidade de compartilhamento de informações promovida pela adoção de plataformas digitais intensificou severamente a extensão deste problema. No entanto, na maioria dos casos, essas plataformas são utilizadas apenas como veículos para disseminação de conteúdos que na verdade são produzidos por websites externos focados especificamente na produção e disseminação de notícias falsas. Assim, neste trabalho apresentamos uma ampla caracterização de websites de baixa e alta credibilidade. Nossos resultados revelam características interessantes dos websites de baixa credibilidade que podem ser úteis, inclusive, para distingui-los dos demais, contribuindo para o entendimento do fenômeno da desinformação no contexto brasileiro.*

1. Introdução

O problema da desinformação tem afetado diversos países, ameaçando a integridade do discurso público, de processos eleitorais e da governança democrática como um todo. No contexto da pandemia do Coronavírus, o problema alcançou novas magnitudes, tornando-se também uma preocupação de saúde pública seja na forma da promoção de remédios sem eficácia comprovada ou da ridicularização de medidas sanitárias (i.e.,uso de máscaras e distanciamento social), podendo assim ocasionar em grandes perdas de vidas [Galhardi et al. 2020]. O problema é reconhecido pelas principais autoridades no Brasil e no mundo: em 2021, o prêmio Nobel da Paz foi concedido a dois jornalistas reconhecidos internacionalmente por seu empenho na frente de combate à desinformação e na defesa da liberdade de expressão[1].

[1]https://www.anj.org.br/premio-nobel-reconhece-a-importancia-do-jornalismo-no-combate-a-desinformacao/

No Brasil, as eleições presidenciais de 2018 foram marcadas pela extensa distorção da verdade e por abusos promovidos por campanhas de desinformação lançadas por entidades com agendas bem definidas, sobretudo em plataformas digitais como o Whatsapp [Melo et al. 2021]. Nesse contexto, a desinformação ganhou uma dimensão sem precedentes no Brasil, potencializada em todos os instantes por plataformas digitais. Hoje, existe um generalizado receio de que o episódio venha a se repetir no contexto das eleições presidenciais de 2022, dessa vez no aplicativo de troca de mensagens Telegram, promovido como uma plataforma digital alternativa que preza pela "discussão livre". No entanto, esta plataforma tem sido alvo frequente de campanhas disseminando informações falsas e discurso de ódio, já tendo sido identificado como a nova fronteira de mobilizações políticas [Júnior et al. 2021].

Apesar da complexidade que aplicativos de troca de mensagens (e.g., Telegram, Whatsapp) trazem para o problema, na maioria das vezes essas plataformas são utilizadas apenas como mecanismos de disseminação de conteúdos de desinformação que na verdade estão veiculados em *websites* externos especializados na difusão de notícias falsas no contexto de campanhas de desinformação específicas. Estes *websites* externos são, portanto, frequentemente a verdadeira raiz da desinformação. Aqui, é notável o fato de que apesar da ampla difusão e popularidade destes *websites* externos em redes sociais e aplicativos de trocas de mensagens, eles têm sido amplamente negligenciados pelos estudos acadêmicos da área de desinformação. Assim, existe uma lacuna a ser preenchida: o entendimento das características desses *websites* como sua localização, aspectos de registro e certificação, entre outros. Isso é de suma importância para o entendimento do ecossistema de desinformação no Brasil, e nesta pesquisa damos um primeiro passo nesta direção.

Neste trabalho, construímos uma coleção de dados composta por *websites* de baixa credibilidade e *websites* de alta credibilidade. Para a construção do conjunto de *websites* de baixa credibilidade, foram coletados *websites* que possuem ao menos uma notícia publicada cuja veracidade é contestada por alguma uma verificação de fatos publicada em uma agência de verificação de fatos signatária da IFCN (*International Fact-Checking Network*[2]). Já o conjunto de *websites* de alta credibilidade foi construído a partir da lista de portais de notícias reconhecidos pela Associação Nacional de Jornais (ANJ)[3]. Assim, depois de construirmos os conjuntos de dados, caracterizamos uma série de aspectos desses *websites*, como propriedades de seu registro DNS, aspectos de sua certificação TLS e também de sua localização geográfica visando entender as diferenças entres os dois conjuntos de credibilidade. Entre os resultados obtidos, destacamos que sites de baixa credibilidade tem vida útil mais curta, utilizam certificados TLS de menor duração e são hospedados no exterior com maior frequência.

O restante do trabalho está organizado da seguinte forma. Na Seção 2 discutimos trabalhos relacionados. Na Seção 3 descrevemos o processo de criação da base de dados de sites de baixa e alta credibilidade. Na Seção 4 oferecemos uma visão geral dos atributos explorados. Em seguida, na Seção 5 detalhamos os resultados obtidos, apresentando a caracterização das principais diferenças entres os *websites* de alta e baixa credibilidade. Por fim, a Seção 6 conclui este trabalho.

[2]https://www.poynter.org/ifcn/

[3]https://www.anj.org.br/associados/

2. Trabalhos Relacionados

É crescente o interesse de pesquisadores de diversas áreas pelo fenômeno da desinformação em plataformas digitais. Em computação, tem havido um grande esforço da comunidade com o objetivo de entender e propor soluções que sejam efetivas contra o problema [Pereira and Marques-Neto 2021, Santos et al. 2019]. Por exemplo, com foco em prover um melhor entendimento do ecossistema de desinformação, [Richard Fletcher and Nielsen. 2018] analisaram a disseminação e alcance de instâncias de desinformação na França, revelando que o *website* de desinformação de maior popularidade havia alcançados 1,5 milhões de franceses. Por outro lado, recentemente existem esforços com foco na construção de ferramentas (i.e., sistemas *Web*) objetivando oferecer transparência relacionada à fenômenos pertinentes ao ecossistema da desinformação [Melo et al. 2019].

Por fim, mais relacionado ao nosso trabalho, existem esforços recentes que buscam criar modelos para diferenciar desinformação de informação. Os trabalhos nessa direção comumente fazem uso de estratégias de aprendizado de máquina e modelagens estatísticas para modelar e criar classificadores capazes de identificar a desinformação a partir de padrões textuais extraídos do conteúdo [Yimin Chen and Conroy. 2015, Nicole O'Brien and Boix. 2018, Martin Potthast and Stein. 2016], ou ainda, informações contextuais extraídas da propagação de conteúdo em mídias sociais [Niall J. Conroy and Chen. 2015, Srijan Kumar and Leskovec. 2016].

Apesar da inegável importância destes trabalhos, eles são limitados. A maioria deles explora o problema de combater a desinformação com foco em instâncias específicas (e.g., uma notícia que já foi checada por uma agência de checagem de fatos), ou usando atributos, que as vezes, só estarão disponíveis para análise depois que uma notícia já se propagou em uma plataforma digital (e.g., número de compartilhamos de uma notícia). Assim, neste trabalho, nosso objetivo é complementar aos trabalhos anteriores que proveem entendimento do problema, oferecendo o diferencial de prover uma caracterização de atributos de redes relacionados a *websites* de baixa credibilidade no Brasil, que, até onde sabemos, foram pouco explorados em esforços anteriores. Atributos de rede não são restritos à instâncias de desinformação específicas e estão disponíveis tão cedo quanto a criação de um novo *website*. Assim, acreditamos que uma caracterização dos padrões associados a esses atributos pode ser útil para a construção de mecanismos que sejam efetivos na identificação precoce de *websites* produtores de desinformação no Brasil.

3. Construção da Base de Dados

Para caracterizar *websites* de baixa credibilidade no Brasil é importante obter um conjunto de dados de alta credibilidade [Reis et al. 2019] que tenha sido rotulado por jornalistas especialistas. Assim, nesta seção são apresentados detalhes relativos ao processo de construção desta base de dados, incluindo a estratégia utilizada para construção de um repositório de *websites* de alta credibilidade que foi explorado para fins de comparação dos resultados.

3.1. Websites de Baixa Credibilidade

Para construção da base de *websites* de baixa credibilidade, foi desenvolvida uma estratégia cujo objetivo era identificar *websites* conhecidos ou emergentes que espalharam

(ou espalham) notícias falsas. Essa abordagem foi desenvolvida com base na hipótese de que uma pessoa (ou até mesmo um robô) capaz de postar uma notícia reconhecidamente falsa em uma plataforma digital (e.g., Twitter) tem capacidade de fazê-lo mais de uma vez.

Primeiramente, houve um esforço manual de identificação de uma notícia falsa que pudesse ser utilizada como raiz para o processo que será descrito a seguir. Para a notícia em questão, era necessário que houvesse uma avaliação da veracidade de seu conteúdo disponibilizada por uma agência de checagem de fatos brasileira, signatária da IFCN (*International Fact-Checking Network*[4]). Durante esta etapa foi selecionada como raiz uma notícia que associava o coronavírus a uma arma biológica[5]. Essa notícia foi desmentida pela agência Estadão Verifica na checagem de fatos disponível em: https://politica.estadao.com.br/blogs/estadao-verifica/coronavirus-site-distorce-entrevista-para-sugerir-possibilidade-de-criacao-de-arma-biologica/

Depois disso, foi implementado um coletor de dados do Twitter[6] que recuperou todos os usuários da referida plataforma que postaram a notícia selecionada como raiz. A partir disso, para todos esses usuários recuperados, foi efetuada a coleta de todas as postagens efetuadas por eles na plataforma e filtradas todas as postagens com links. Todos os links foram incluídos em uma lista, e utilizados como base para uma nova execução do processo – esse procedimento poderia ser repetido indefinidamente, até que todos websites pertencentes a esse componente conexo fossem alcançados, todavia permite haja uma interrupção quando um máximo m de websites suspeitos é alcançado, evitando que a inspeção manual a posteriori deixe o tempo de execução impraticável. Ao final do processo, foi efetuada uma extração dos *websites* associados a cada um dos links para a construção de uma lista de potenciais *websites* de baixa credibilidade. Para cada *website* desta lista, buscamos no Twitter os 20 links mais compartilhados que foram publicados pelo referido *website*.

Em seguida, para cada link recuperado foi efetuada a extração do título da notícia, e verificada a correspondência com títulos de checagens de fatos realizadas e disponibilizadas por agências brasileiras. O objetivo desta etapa é verificar que a notícia publicada continha alguma desinformação que já havia sido desmentida por uma agência. Os dados de checagens explorados durante este processo, foram coletados e disponibilizados em trabalhos anteriores ([Couto et al. 2021]). Como resultado, para cada link, foi computado um grau de similaridade entre o seu título e o título de cada uma das checagens de fatos presentes no repositório utilizado como base. Para cada instância (i.e., link) foi selecionada a checagem com maior grau similaridade e efetuada uma inspeção manual cujo objetivo era garantir que ambos (notícia e checagem) realmente tratavam do mesmo assunto. Em outras palavras, foi verificado se o link realmente continha uma notícia falsa que foi desmentida pela referida checagem. Para todos os casos em que isso foi verdade, o referido *website* foi incluído em uma base de dados contendo sites de baixa credibilidade. Essa base de dados, composta por 41 *websites* distintos, foi utilizada para realização da

[4]A IFCN, disponível em https://www.poynter.org/ifcn/ é uma rede internacional de checagem de fatos que estabelece diretrizes e boas práticas para o processo de checagem de fatos.

[5]https://www.estudosnacionais.com/32878/novo-coronavirus-pode-ter-sido-fabricado-como-arma-biologica-afirma-pesquisadora/

[6]O Twitter foi selecionado considerando que esta é uma plataforma bastante utilizada no Brasil.

caracterização proposta neste trabalho.

3.2. Websites de Alta Credibilidade

Para fins de comparação dos resultados de caracterização e investigação do potencial dos atributos de rede para distinguir um *website* de baixa credibilidade dos demais, foi construída uma segunda base de dados contendo *websites* de alta credibilidade. Para construção desta base de dados, foi utilizada como base a lista de portais de noticias digitais credenciados pela Associação Nacional de Jornais (ANJ)[7]. A ANJ é uma entidade sem fins lucrativos reconhecida internacionalmente por observar os princípios da responsabilidade sobretudo no contexto do combate à desinformação. O estatuto da ANJ [8] prevê a obediência obrigatória por parte de todos os seus associados do código de ética da instituição, que defende a apuração e publicação da verdade dos fatos de interesse público, não admitindo que sobre eles prevaleçam quaisquer interesses. No total, foram coletados 98 *websites* desta lista para compor a nossa base de dados de alta credibilidade.

4. Fontes de Atributos

A implantação e operação de um *website* na Internet requer obtenção de recursos bem como configuração de protocolos e arcabouços. Os recursos podem variar e ter viés diferente em função de custo, procedência, duração, entre outros fatores. De forma similar, configurações podem variar e ter viés em função da carga de trabalho e experiência do time de técnicos. Neste trabalho utilizamos atributos relacionados aos recursos e configurações utilizados por um *website* para relacioná-los com a credibilidade do conteúdo hospedado. A análise do potencial de atributos de rede para a detecção de *websites* de baixa credibilidade envolveu a implementação de extratores para 31 atributos de três fontes de dados:

- **Domínio:** Atributos relacionados com o registro, operação e configuração do nome de domínio, incluindo dados no DNS.
- **Certificado:** Atributos sobre aspectos de segurança do domínio e *website* extraídos a partir de atributos do certificado TLS (ou ausência dele).
- **Geolocalização:** Atributos obtidos a partir da geolocalizaçao do endereço IP hospedando um *website*.

A Tabela 1 apresenta uma visão geral dos atributos computados. Os atributos computados de todas as fontes estão disponíveis no momento da criação de um *website*, permitindo a criação de sistemas de *flagging* com baixo tempo de resposta. Além disso, os atributos podem ser obtidos através de serviços públicos (e.g., WHOIS[9]) ou através de ferramentas comerciais de baixo custo (e.g., IPStack, utilizado aqui para fazer a geolocalização dos diferentes *websites*).

Apesar disso, a extração de todos os atributos de todos os *websites* contidos na base dados provou-se uma tarefa desafiadora, em particular pela falta de padronização na resposta dos diferentes protocolos utilizados. Os dados presentes em uma resposta a uma consulta WHOIS, por exemplo, diferem significantemente entre registros de domínio

[7] https://www.anj.org.br/associados/

[8] https://www.anj.org.br/wp-content/uploads/2021/04/ANJ_ESTATUTO_SOCIAL.pdf

[9] Protocolo de consulta a informações de registro associados com entidades na Internet.

Identificador	Tipo	Descrição
Atributos de domínio		
subdomain-hifen	Bool.	URL contém um hífen
subdomain-digit	Bool.	URL contém um dígito
tld-br-or-com	Bool.	TLD (Top Level Domain) é .br ou .com
news-keywords	Bool.	URL contém uma keyword indicativa de um portal de notícias (e.g., ”gazeta ”jornal”tribuna”)
whois-privacy	Bool.	O registrante utiliza opções de privacidade para queries WHOIS
resolution-hops	Num.	Número de 'hops' necessários para resolução do subdomínio em um endereço IP
caa-txt-count	Num.	Número de entradas do tipo CAA ou TXT no registro DNS do domínio
len-subdomain	Num.	Número de careteres no subdominio do portal
domain-age	Num.	Tempo, em dias, desde o registro inicial do domínio
domain-expiry	Num.	Tempo, em dias, até expiração do domínio
domain-update	Num.	Tempo, em dias, desde a última modificação no registro DNS do domínio
as-n	Cat.	Autonomous System Number associado com o IP para qual o subdominio foi resolvido
registrar	Cat.	Entidade responsável pelo registro do domínio (e.g., GoDaddy)
registrar-url	Cat.	URL da entidade responsável pelo registro do domínio (e.g., NameCheap.com)
Atributos de Certificado		
allows-http	Bool.	Servidor retorna conteúdo em requisições HTTP
redirects-http	Bool.	Servidor redireciona requisições
ca-is-letsencrypt	Bool.	Emissor de certificado TLS é o popular serviço de certificação ”Let's Encrypt”
cert-expired	Bool.	TLS encontrado está expirado
public-key-bits	Num.	Número de bits utilizados na chave pública durante handshake TLS
cert-age	Num.	Tempo, em dias, desde a emissão do certificado TLS
cert-expiry	Num.	Tempo, em dias, até expiraçao do certificado TLS
cert-lifespan	Num.	Tempo de validade total do certificado TLS (emissão até expiração)
ca-entity	Cat.	Entidade emissora do certificado TLS
ca-nacionality	Cat.	País associado com a entidade emmissora do certificado TLS
Atributos de geolocalização		
ip-in-brazil	Bool.	Geolocalizaçao do endereço IP resultou em coordenadas no Brasil
ip-in-usa	Bool.	Geolocalizaçao do endereço IP resultou em coordenadas no Estados Unidos da América
as-ip-equal-cc	Bool.	Geolocalizaçao do endereço IP resultou no mesmo país em que o ASN está registrado
ip-cc	Cat.	Código de país resultado da geolocalizaçao do endereço IP
as-cc	Cat.	Código de país associado com o registro da ASN
lat-ip	Num.	Latitude resultante da geolocalizaçao do endereço IP
long-ip	Num.	Longitude resultante da geolocalizaçao do endereço IP

Tabela 1. Atributos extraídos separados por categoria.

(*registrars*) e em particular entre sites registrados no *top-level domain* .br e .com. Dessa forma, uma normalização foi necessária: consideramos apenas atributos calculáveis a partir das informações presentes nas respostas de todos os sites em nossa base de dados. *Websites* que se tornaram inativos entre a identificação dos conjuntos de sites e a coleta dos atributos, ou seja, que deixaram de retornar conteúdo para requisições HTTP/HTTPS, foram eliminados da base de dados.

4.1. Atributos de Domínio

A resolução de nomes via o sistema DNS (*Domain Name Service*) é essencial para o funcionamento de qualquer *website* público. Nesse sentido, o registro de domínio e a configuração dos seus servidores DNS autoritativos oferecem informações essenciais sobre a natureza do mesmo, possibilitando o mapeamento de padrões que servem de indica-

tivos não apenas da robustez infraestrutural de um site como possivelmente a intenção do registrante no momento em que o domínio foi criado.

Um exemplo intuitivo dessa noção é a duração do registro de um nome e o período até sua data de expiração. Ao iniciarem novas campanhas de desinformação, é razoável esperar que registrantes tenham ciência da possibilidade que em algum momento o *website* associado venha a ser derrubado por ordem judicial e que, portanto, este tipo de *website* tenha uma têndencia maior a efetivar renovações de registro mais curtas, evitando maiores perdas financeiras. De fato, como veremos na seção de resultados, os valores observados de dias até a expiração de *websites* de baixa credibilidade são bem menores que os observados em sites de alta credibilidade. De forma similar, outros atributos associados com o domínio, como o uso de opções de privacidade no protocolo WHOIS ou a natureza da entidade registradora utilizada para efetivar a criação do domínio, também são indícios que podem estar correlacionados com a intenção do registrante e uso de um determinado domínio.

4.2. Atributos de Certificado

O suporte ao acesso encriptado a *websites* públicos está em crescente adoção na Internet, particularmente através do uso do protocolo *HyperText Transfer Protocol Secure* (HTTPS) onde antes utilizava-se *HyperText Transfer Protocol* (HTTP). A adoção o uso de HTTPS por um site envolve a emissão de certificados TLS, necessários para autenticar a identidade de servidores respondendo por requisições direcionadas a um domínio. Nesse contexto, de forma similar ao DNS, podemos extrair atributos possivelmente correlacionadas com a natureza dos diferentes domínios analisados. Em particular, a maioria dos serviços de geração e manutenção de certificados TLS ocasionam custos recorrentes, frequentemente proporcionais ao suporte, à robustez, ou propriedades do serviço contratado.

Esse fenômeno se manifesta de diferentes maneiras. Para fins de ilustração, destacamos o período de validade total do certificado: serviços grátis de certificação TLS geralmente emitem certificados com validade menor ou igual a 90 dias enquanto serviços pagos com frequência emitem certificados com duração de um ano ou mais. Nesse cenário, por motivos análogos aos apresentados no contexto do registro DNS de um domínio da subseção anterior, isto é, evitar maiores perdas financeiras, é razoável esperar que sites de baixa credibilidade façam, com maior probabilidade, uso de serviços de baixo (ou nenhum) custo. Veremos que em nossa base de dados a maioria dos certificados dos sites de baixa credibilidade expiram em até 90 dias, então a maioria dos de alta credibilidade expiram apenas um ano após a coleta dos dados.

Outras propriedades de certificados podem estar relacionadas à natureza dos domínios. O número de bits presentes na chave pública empregada durante o *handshake* (verificação de identidades, negociação de protocolos de cifragem e troca de chave simétrica) entre clientes e o servidor de um determinado portal de notícias é outro indicador da robustez do serviço de certificação empregado por um *website*.

4.3. Atributos de Geolocalização

No Brasil, recorrentes iniciativas governamentais objetivam abater vetores de notícias falsas[10]. Iniciativas desse tipo têm fácil acesso a uma gama de ferramentas jurídicas que

[10] https://bit.ly/3Hs4bgY

podem ser acionadas para efetivar o rápido desligamento de *websites* ou páginas em redes sociais. Medidas para dificultar esse processo passam então a ser do interesse de entidades que buscam lançar campanhas de desinformação. Nesse contexto, a utilização de serviços de registro e hospedagem de domínios oferecidos por entidades estrangeiras representam uma das mais populares medidas deste tipo. Ao efetivamente operarem sob a jurisdição de outro país, *websites* de baixa credibilidade aumentam significantemente a complexidade jurídica de processos objetivando seu desligamento. Nesse contexto, a geolocalização de um *website* se torna um objeto de interesse que pode conter informações úteis para diferenciar fontes de informação de alta e baixa credibilidade.

5. Resultados

Nesta seção apresentamos os resultados obtidos a partir dos atributos das três fontes de dados discutidas na Seção 4. Aqui, temos como objetivo caracterizar e investigar a capacidade discriminativa dos atributos extraídos. Para tal, procuramos evidenciar que os atributos calculados sobre o conjunto de *websites* de baixa credibilidade seguem distribuições fundamentalmente diferentes daquelas observadas para o conjunto de alta credibilidade.

Em particular, para os atributos numéricos (identificados na Tabela 1) os resultados são apresentados a partir de três medidas principais: média, 60º percentil, e o p-valor do teste de Kolmogorov-Smirnov [Massey Jr 1951]. Este teste metrifica a distância entre duas distribuições de probabilidade acumulada. Em nosso caso, as duas distribuições são relativas ao conjunto de *websites* de alta e baixa credibilidade. Em essência, o teste responde à pergunta "*qual a probabilidade que os dois conjuntos de valores vieram de uma mesma população?*" Em outras palavras, o teste nos permite dizer que um certo atributo tem valores fundamentalmente diferentes entre os sites de baixa e alta credibilidade. Se o p-valor associado com um teste é inferior a um nível de significância (utilizaremos 0.05, ou, 5%) então esperamos que esse atributo possa contribuir para diferenciar entre sites de alta e baixa credibilidade (e.g., podem servir de entrada para um classificador baseado em aprendizado de máquina). Para os atributos categóricos e booleanos, apresentaremos a incidência, em pontos percentuais, de cada categoria (falso ou verdadeiro no caso dos atributos booleanos) observada nas duas populações. Novamente, atributos com incidência muito diferentes podem ser úteis para diferenciar entre sites de alta e baixa credibilidade.

Nas subseções seguintes, apresentamos, para cada fonte de atributos, os resultados dos dois atributos numéricos com distribuições mais dissimilares, a incidência das classes de atributos categóricos e booleanos, bem como observações sobre os demais atributos.

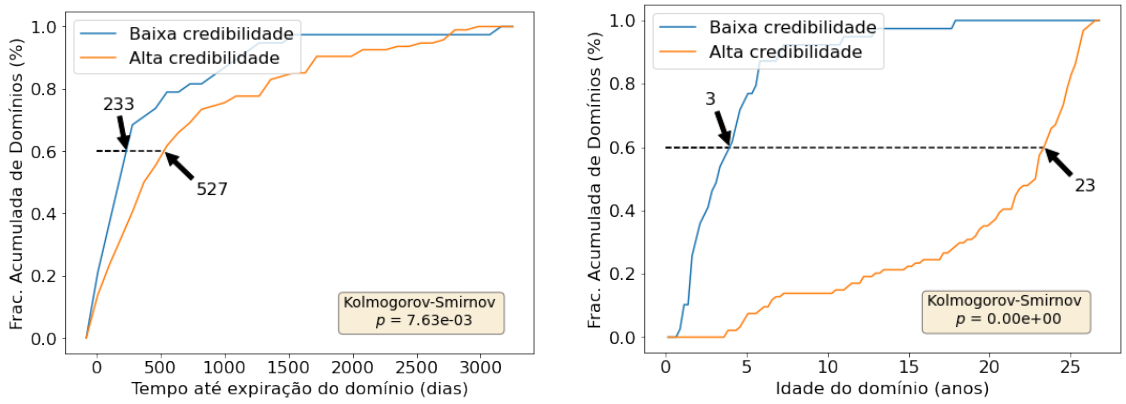
5.1. Domínios

Obtemos os valores de atributos do registro de um nome de domínio (e.g., registrante, idade e prazo de expiração) fazendo consultas a servidores WHOIS. Obtemos valores de atributos da configuração do servidor autoritativo de um nome de domínio (e.g., número de entradas CAA/TXT) realizando consultas DNS através do comando `dig`. Em ambos os casos, os resultados das consultas são filtrados para extração de campos de interesse através de *scripts* Python que levam em consideração os diferentes formatos de respostas para essas consultas, que variam em função do TLD sob o qual o domínio foi registrado.

Tempo até expiração do domínio. O tempo de expiração de um domínio é função do período de registro pago antecipadamente. Neste trabalho investigamos se sites de baixa

credibilidade contratam o serviços de registro por menor duração quando comparados à sites de alta credibilidade.

A Figura 1a corrobora essa propriedade: 60% dos *websites* de baixa credibilidade expira em até 233 dias após a coleta dos dados enquanto o mesmo valor para o *websites* de alta credibilidade é de 527 dias. O p-valor calculado no teste de Kolgomorov-Smirnov evidencia que as distribuições dos tempos de expiração são diferentes. Conclui-se que o tempo até a data expiração de um domínio pode ser utilizada como um fator contribuinte na caracterização de *websites* de baixa credibilidade.



(a) Tempo até expiração dos domínios

(b) Idade dos domínios

Figura 1. CDF dos atributos numéricos de domínio com menor similaridade.

Idade do domínio. A idade de um domínio pode ser vista como um indicador de histórico e consistência do empenho em sua manutenção e criação de conteúdo a médio e longo prazo. Aqui, buscamos averiguar se *websites* de baixa credibilidade tendem a ser estabelecidos há menos tempo por terem uma vida útil mais curta possivelmente por buscarem atender agendas efêmeras ou serem derrubados por ordem judicial com maior frequência.

A Figura 1b dá forte suporte a este argumento: 76% dos sites de baixa credibilidade têm menos de 5 anos desde seu registro. *Websites* de alta credibilidade apresentam uma vida útil muito mais longa: apenas 6,7% deles têm menos de 5 anos desde seu registro. O teste de Kolgomorov-Smirnov resultou em um p-valor arredondado para zero, indicando que este atributo se comporta de maneira diferente entre os dois conjuntos de *websites* e portanto pode ser útil para distingui-los.

Atributos de domínio categóricos e booleanos. Na Tabela 2 podemos observar o grande discrepância do atributo categórico ip-cc (país de geolocalização do IP): apenas 12% dos sites de baixa credibilidade estão hospedados no Brasil, enquanto a incidência entre os sites de alta credibilidade é 48,4%. Este resultado indica que *websites* de baixa credibilidade operaram no exterior com frequência significativamente maior quando comparados a *websites* de alta credibilidade. Entre os atributos booleanos apresentados na tabela 3, três se destacam: a presença de hífen no subdomínio (9,8% vs 3,1%), a presença de dígitos no domínio (17,1% vs 7,1%) e a presença de palavras-chave jornalísticas no subdomínio (24,4% vs 50%). Ademais, a diferença de incidência no atributo tld-br-or-com indica que *websites* de baixa credibilidade podem ser mais propícios a utilizarem TLDs incomuns, enquanto 98% dos sites de alta credibilidade estão registrados sob os TLDs .com e .br.

		Código de país associado							
		US	**BR**	**CN**	**SG**	**GB**	**CA**	**BE**	**DK**
IP	Alta credibilidade (%)	48,4	48,4	0,0	2,1	0,0	1,1	0,0	0,0
	Baixa credibilidade (%)	82,1	12,8	0,0	2,6	0,0	0,0	0,0	2,6
ASN	Alta credibilidade (%)	56,8	42,1	0,0	0,0	0,0	1,1	0,0	0,0
	Baixa credibilidade (%)	84,6	10,3	0,0	0,0	0,0	2,6	0,0	2,6
Emissor TLS	Alta credibilidade (%)	85,7	2,4	1,2	0,0	3,6	0,0	7,1	0,0
	Baixa credibilidade (%)	94.4	0,0	0,0	0,0	5,6	0,0	0,0	0,0

Tabela 2. Distribuição de classes dos atributos categóricos.

		Incidência de atributos booleanos	
		Baixa credibilidade (%)	**Alta credibilidade (%)**
Domínio	subdomain-hifen	9,8	3,1
	subdomain-digit	17,1	7,1
	tld-br-or-com	85,4	98,0
	news-keywords	24,4	50,0
	whois-privacy	7,3	5,1
Certificado	allows-http	12,2	14,3
	redirects-http	80,5	78,6
	ca-is-letsencrypt	31,7	32,7
	cert-expired	12,2	16,3
Geolocalização	ip-in-brazil	12,8	48,4
	ip-in-usa	82,1	48,4
	as-ip-equal-cc	95,1	91,8

Tabela 3. Incidência dos atributos booleanos por grupo de credibilidade.

5.2. Certificados

Com o objetivo de verificar as intuições acerca da credibilidade de *websites* quantificadas por atributos de certificados TLS propostas na Seção 4.2, extraímos atributos de certificado utilizando comandos da biblioteca OpenSSL[11] e também da ferramenta de linha de comando cURL.[12] Em particular, o OpenSSL foi utilizado para a extração de atributos associados com os certificados, como as datas de emissão e expiração dos certificados TLS, e o cURL foi utilizado para determinar se um *website* aceita requisições HTTP e se estas conexões são automaticamente redirecionadas para uma requisição HTTPS.

Duração dos certificados. A duração de um certificado TLS associado com um *website* é função do tipo de serviço de certificação TLS contratado pelos *websites*. Neste contexto, investigamos se certificados emitidos para sites de alta credibilidade tem um período de validade maior do que aqueles emitidos para *websites* de baixa credibilidade. Observamos que todos os certificados tinham duração múltimo de 3 meses, resultando em uma quantidade pequena de valores possíveis e tornando mais adequada sua apresentação na forma de uma tabela de incidência.

A Tabela 4 mostra a distribuição das diferentes durações dos certificados na base

[11] https://www.openssl.org/

[12] https://curl.se/

		Tempo de vida (em dias)		
		90	**364**	**>364**
Certificados	Alta credibilidade (%)	41,7	34,5	23,9
	Baixa credibilidade (%)	52,8	44,9	2,4

Tabela 4. Distribuição da vida útil de certificados.

de dados. Observamos que enquanto 23,9% dos certificados de sites de alta credibilidade tem duração superior a um ano, isso acontece para apenas 2,4% dos sites de baixa credibilidade. Este resultado indica que certificados de maior duração, em particular com duração superior a um ano, constituem um indício de credibilidade. Neste contexto, notamos que o Let's Encrypt, um dos maiores serviços grátis para emissão de certificados TLS emite apenas certificados com duração de 3 meses[13].

Atributos categóricos e booleanos. A Tabela 3 oferece a incidência de cada uma das features booleanas nos dois conjuntos de credibilidade. Aqui, observamos que nenhum atributo demonstrou se comportar de maneira diferentes entre os dois grupos em vista da incidência praticamente igual em todos os atributos. Assim, os dados apontam que estes, assim como a maioria dos atributos de certificado, não contribuem para a determinação de credibilidade de um *website*.

5.3. Geolocalização

Utilizamos a API de geolocalização IPStack (https://ipstack.com) para obter as coordenadas geográficas associadas ao endereço IP para qual cada domínio foi resolvido. A Figura 2 evidencia de forma intuitiva a existência de uma correlação entre as coordenadas obtidas e a publicação de conteúdo de desinformação. É fácil observar que as coordenadas do conjunto de sites de baixa credibilidade está mais geograficamente distribuída e mais intensamente presente no exterior, como veremos nas análises seguintes.

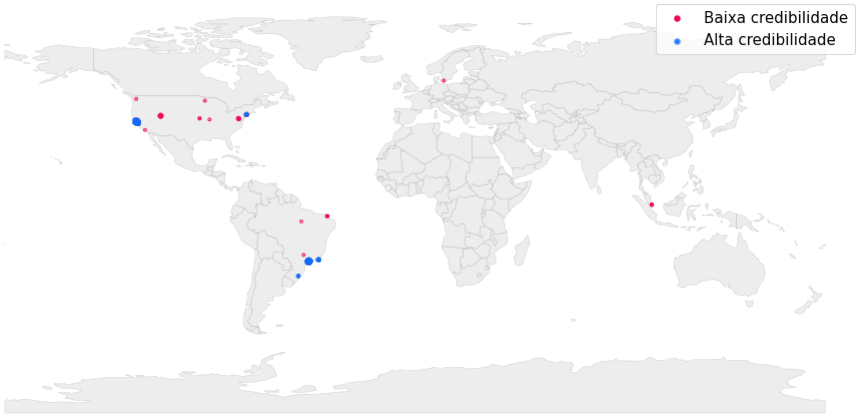
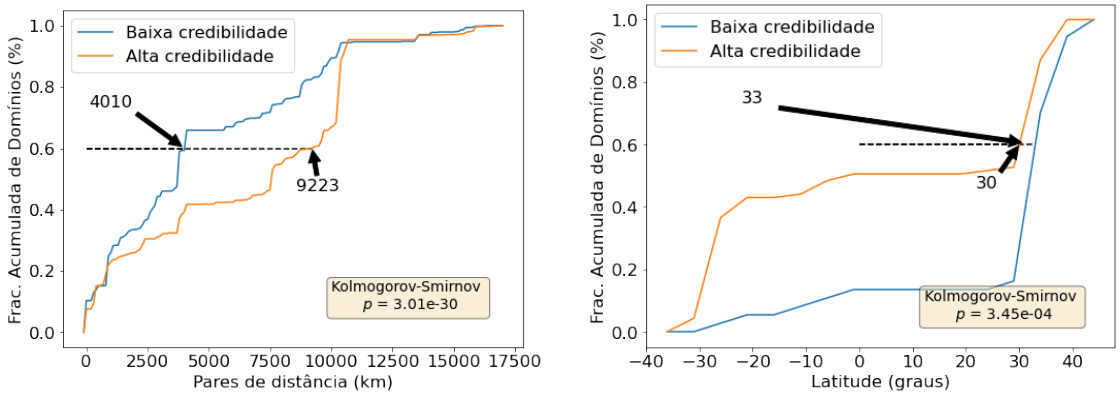


Figura 2. Mapa com as coordenadas de *websites* por conjunto de credibilidade.

Concentração geográfica. Muitos dos principais serviços de hospedagem de *websites* se encontram em polos tecnológicos. O mapa da Figura 2 sugere que sites de baixa

[13] https://letsencrypt.org/2015/11/09/why-90-days.html



(a) Pares de distâncias por grupo de credibilidade

(b) Latitude dos domínios

Figura 3. CDF dos atributos numéricos de geolocalização de menor similaridade.

credibilidade tem uma tendência maior a utilizar serviços alternativos fora destes centros. Para averiguar essa hipótese, extraímos as cidades associadas às coordenadas de cada *website*: observamos que os 98 *websites* de alta credibilidade estão hospedados em apenas 25 cidades, enquanto os 41 sites de baixa credibilidade se encontram em 20 cidades, sugerindo que os *websites* de baixa credibilidade de fato se encontram mais espalhados.

Pares de distâncias. A fim de explorar a intuição proposta na Seção 4.3, computamos a distância geográfica entres pares de *websites* em cada grupo de credibilidade. Essa métrica nos permite observar se os *websites* de alta ou baixa credibilidade estão mais ou menos concentrados geograficamente. Para calcular essas distâncias, utilizamos a distância de Vincenty implementada na biblioteca Geopy[14]. Assim, calculamos para cada *website* a distância entre suas coordenadas e as coordenadas de todos os outros *websites* em seu grupo de credibilidade (i.e., baixa ou alta).

A Figura 3a apresenta a distribuição cumulativa (CDF) das distâncias nos dois grupos, apontando o 60º percentil e o p-valor obtido no teste de Kolmogorov–Smirnov. Aqui, é importante observar que os *websites* de baixa credibilidade estão mais intensamente concentrados nos EUA do que os sites de alta credibilidade estão concentrados no Brasil, conforme evidenciado na Tabela 2. Assim, os *websites* de baixa credibilidade geram pares de distâncias menores apesar de estarem presentes em uma quantidade maior de países em volta do globo. Em particular, a moda de aproximadamente 10.000 km entre os *websites* de alta credibilidade captura a distância entre São Paulo e São Francisco; enquanto a moda de aproximadamente 4.000 km entre os *websites* de baixa credibilidade captura a distância entre as costas dos EUA. O p-valor obtido no teste de Kolmogorov–Smirnov entre os dois conjuntos de pares de distância foi muito menor que 5%, confirmando a diferença entre as distribuições.

Latitude. Muitos serviços de computação em nuvem ou hospedagem de conteúdo online estão amplamente disponíveis e acessíveis tanto no Brasil quanto no exterior. A latitude nos permite verificar de forma mais direta em qual hemisfério *websites* estão hospedados. Em particular, grande parte dos centros computacionais (*datacenter*) e serviços de hospedagem no exterior concentram-se nos Estados Unidos e Europa.

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

De fato, a Figura 3b indica que *websites* de baixa credibilidade realmente têm maior probabilidade de estarem hospedados no hemisfério norte. Aqui vale destacar que a latitude zero separa os hemisférios e, portanto, através do mapa da Figura 2 podemos correlacionar latitudes maiores que zero com sites registrados no exterior (os sites hospedados no norte do Brasil foram geolocalizados em Fortaleza-CE e Parauapebas-PA, que estão abaixo do Equador). Nesta direção, a Figura 3b confirma a ideia inicial uma vez que menos de 20% dos *websites* de baixa credibilidade estão hospedados abaixo da linha do equador, enquanto para os sites de alta credibilidade temos aproximadamente 50%.

6. Conclusão

Atualmente existe um grande conjunto de esforços voltados para a desinformação focados em identificar se uma notícia especifica é falsa e negligenciando aspectos dos *websites* onde elas são veiculadas. Assim, neste trabalho apresentamos uma ampla caracterização de *websites* de baixa e alta credibilidade no contexto brasileiro, ressaltando suas diferenças. Neste contexto, foram construídas as bases de dados exploradas nesta pesquisa. Para a construção do conjunto de dados de baixa credibilidade, foram coletados *websites* que possuem ao menos uma notícia cuja veracidade tenha sido contestada por alguma agência de verificação de fatos brasileira. Por lado, para compor o conjunto de dados de *websites* de alta credibilidade foi explorada a lista de portais de notícias reconhecidas pela ANJ. Depois disso, foram computados 31 atributos desses *websites*, incluindo características de domínio, certificado e geolocalização.

Nossos resultados revelam que apenas 12,2% dos *websites* de baixa credibilidade são registrados e hospedados no Brasil (para *websites* alta credibilidade esse percentual é de 47%). Além disso, mostramos que *websites* de alta credibilidade possuem uma vida útil mais longa, além de serem registrados por mais tempo e possuírem certificados TLD validos por períodos de tempo mais longos. Descobrimos ainda, que *websites* de baixa credibilidade são mais propícios a utilizar TLD. Em suma, os resultados apresentados revelam características interessantes dos *websites* de baixa credibilidade, contribuindo para o entendimento do fenômeno da desinformação no contexto brasileiro. Acreditamos que nossas descobertas revelam aspectos em comum desses *websites* que podem ser úteis para distingui-los dos demais, abrindo uma nova avenida de trabalhos futuros que possa ser explorada principalmente pela comunidade de redes de computadores.

Agradecimentos

Este trabalho foi realizado com apoio financeiro do Ministério Público de Minas Gerais (MPMG), projeto Capacidades Analíticas, além de Fapesp, CNPq, e Fapemig.

Referências

Couto, J., Pimenta, B., de Araújo, I. M., Assis, S., Reis, J. C. S., da Silva, A. P., Almeida, J., and Benevenuto, F. (2021). Central de fatos: Um repositório de checagens de fatos. In *Anais do Dataset Showcase Workshop (DSW/SBBD)*.

Galhardi, C. P., Freire, N. P., de Souza Minayo, M. C., and Fagunde, M. C. M. (2020). Fato ou fake? uma análise da desinformação frente à pandemia da covid-19 no brasil. *Ciência & Saúde Coletiva*, 25:4201–4210.

Júnior, M., Melo, P., da Silva, A. P. C., Benevenuto, F., and Almeida, J. (2021). Towards understanding the use of telegram by political groups in brazil. In *Proc. of the Brazilian Symposium on Multimedia and the Web (WebMedia)*.

Martin Potthast, Johannes Kiesel, K. R. J. B. and Stein., B. (2016). A stylometric inquiry into hyperpartisan and fake news. *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78.

Melo, P., Benevenuto, F., Kansaon, D., Mafra, V., and Sá, K. (2021). Monitor de whatsapp: Um sistema para checagem de fatos no combate à desinformação. In *Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)*.

Melo, P., Messias, J., Resende, G., Garimella, K., Almeida, J., and Benevenuto, F. (2019). Whatsapp monitor: A fact-checking system for whatsapp. *Proc. of the Int 'l AAAI Conference on Web and Social Media (ICWSM)*.

Niall J. Conroy, V. L. R. and Chen., Y. (2015). Automatic deception detection: Methods for finding fake news. *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Nicole O'Brien, Sophia Latessa, G. E. and Boix., X. (Nov. 2018). The language of fake-news: Opening the black-box of deep learning based detectors. *Proc.of the Workshop on AI for Social Good. Co-located with the Conference on Neural Information Processing Systems (NIPS)*.

Pereira, C. G. and Marques-Neto, H. T. (2021). Caracterização da reação de agências de fact-checking às publicações sobre a pandemia da covid-19 em redes sociais. In *Anais do Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*.

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

Richard Fletcher, Alessio Cornia, L. G. and Nielsen., R. K. (2018). Measuring the reach of "fake news" and online disinformation in europe. *Tech. rep. Reuters Institution and University of Oxford*.

Santos, W. R., Xavier, M. R., da Cunha, D. C., Júnior, J. C., Adauto, D. A., and Ferraz, C. A. (2019). Trendsbot: Verificando a veracidade das mensagens do telegram utilizando data stream. In *Anais Estendidos do Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*.

Srijan Kumar, R. W. and Leskovec., J. (Apr. 2016). Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. *World Wide Web Conference (WWW)*.

Yimin Chen, V. L. R. and Conroy., N. (2015). Towards automatic fake news detection in digital platforms: properties, limitations, and applications. *Proc. of the European Conference on Principles of Knowledge Discovery and Data Mining (PKDD)*.