Identificando Websites de Desinformação no Brasil

Leandro Araújo*[1], Luiz Felipe Nery*[1], Isadora C. Rodrigues[1], João M. M. Couto[1], Julio C. S. Reis[2], Ana P. C. Silva[1], Jussara M. Almeida[1], Fabrício Benevenuto[1]

[1]Depto. Ciência da Computação – Universidade Federal de Minas Gerais (UFMG) – Brasil

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

{leandroaraujo,luiznery,isadorarodrigues,joaocouto}@dcc.ufmg.br

jreis@ufv.br,{ana.coutosilva,jussara,fabricio}@dcc.ufmg.br

Abstract. *In this work, we propose a methodology to identify websites responsible for creating and disseminating misinformation on digital platforms in the Brazilian context. We apply our approach on Twitter. Preliminary results present evidence on the efficacy of the proposed methodology to identify misinformation websites what can be helpful in understanding this phenomenon and allow public organizations to contain the problem in different contexts.*

Resumo. *Neste trabalho propomos uma metodologia para identificação de websites responsáveis pela produção e disseminação de desinformação em plataformas digitais no contexto brasileiro. Aplicamos nossa abordagem no Twitter e os resultados preliminares apresentam evidências do potencial da metodologia proposta para identificação de websites de desinformação que podem ser úteis para entendimento do fenômeno, bem como para atuação do poder público para contenção do problema em diversos contextos.*

1. Introdução

A desinformação é um problema mundial com impactos significativos em diversas esferas da sociedade, tanto em nível individual quanto global (e.g., políticas públicas, processos democráticos) [Galhardi et al. 2020, Allcott and Gentzkow 2017]. Neste contexto, plataformas digitais como Twitter, Facebook e WhatsApp emergiram como meios amplamente utilizados por campanhas de desinformação. Através dessas plataformas, tanto os próprios produtores de conteúdo quanto outros usuários podem ativamente contribuir para a difusão de desinformação em diferentes contextos, como saúde e política [Bessi and Ferrara 2016, Roozenbeek et al. 2020, Martins et al. 2021].

Um elemento central em qualquer ecossistema de desinformação no ambiente online são os *websites* que, com motivação política e/ou financeira, atuam na produção e disseminação de conteúdos de veracidade contestável [Bozarth and Budak 2021]. Assim, é extremamente importante a criação de mecanismos que sejam capazes de identificar tais *websites* em diferentes contextos. Como resultado, esses mecanismos podem prover uma lista de *websites* responsáveis por produzir e disseminar desinformação em plataformas digitais com potencial de ser utilizada para o entendimento de diferentes fenômenos nestes ambientes, bem como para a proposição de medidas que sejam efetivas para a contenção do problema.

*Estes autores contribuíram igualmente para realização deste trabalho.

Apesar da inegável importância desta tarefa, não é trivial encontrarmos listas com *websites* classificados como produtores e/ou disseminadores de desinformação. Em parte, essa dificuldade pode ser justificada pelo fato de que campanhas de desinformação são muitas vezes promovidas e articuladas por organizações poderosas: qualquer indivíduo que se proponha a listar esses *websites* torna-se vulnerável a ações de intimidação, sejam elas realizadas por milícias digitais[1] ou através de assédio jurídico – que consiste na utilização do poder judiciário como forma de perseguição e intimidação[2].

Assim, neste trabalho propomos uma metodologia para identificação de *websites* responsáveis pela produção e disseminação de desinformação em plataformas digitais. A metodologia descrita pode ser executada de maneira independente pelos interessados no estudo desse fenômeno, permitindo que entidades e/ou organizações competentes possam encontrar *websites* dedicados a espalhar tal tipo de conteúdo e atuar, em múltiplos contextos, na contenção do problema. Mais especificamente, aqui focamos no cenário brasileiro, aplicando a metologia proposta na plataforma Twitter. Nossos resultados mostram o potencial da abordagem proposta para a criação de uma lista de *websites* de desinformação.

O restante do trabalho está organizado conforme detalhado a seguir. Na Seção 2 é apresentada a metodologia proposta para identificação de *websites* de desinformação. Em seguida, na Seção 3, são apresentados detalhes relativos à aplicação da metodologia no Twitter. Os resultados preliminares obtidos são discutidos na Seção 4. Por fim, a Seção 5 conclui este trabalho destacando contribuições e direções para trabalhos futuros.

2. Metodologia para Identificação de *Websites* de Desinformação

Neste trabalho, o termo "desinformação" é usado em referência a uma notícia, propagada por alguma mídia digital, contendo informações falsas [Sharma et al. 2019]. Diante disso, um *website* é considerado como de desinformação (ou de baixa credibilidade) se existe pelo menos uma agência de checagem de fatos com especialistas (i.e., jornalistas) que tenha classificado qualquer conteúdo produzido ou reproduzido pelo mesmo como sendo duvidoso, questionável ou de baixa credibilidade (e.g., *fake news*). Assim, com objetivo de fomentar políticas de combate à desinformação em plataformas digitais, nesta seção detalhamos a metodologia proposta para a identificação de *websites* que atuam frequentemente na produção e disseminação de tal tipo de conteúdo. Nesse sentido, é válido ressaltar que a metodologia proposta é baseada na interação dos usuários em plataformas digitais. Mais especificamente, a premissa principal na qual a nossa metodologia se baseia é que se um usuário, em algum momento, compartilha um conteúdo de desinformação, esse usuário tende a compartilhar outros conteúdos de desinformação em outras ocasiões. Além disso, é importante mencionar que a descoberta e/ou identificação destes *websites* de desinformação não é uma tarefa trivial, uma vez que não se restringe somente à coleta de informações compartilhadas por eles, mas também à uma curadoria dessas informações para que uma classificação confiável do conteúdo seja realizada.

A Figura 1 apresenta uma visão geral da metodologia proposta que pode ser dividida em 5 etapas principais. A partir de *links* (pertencentes a um *website*) para notícias que contenham desinformação (i.e., sementes), ou seja, notícias que foram verificadas como enganosas por uma agência de checagem de fatos, é realizada a identificação de

[1]https://www.latimes.com/91910540-132.html

[2]https://www.abraji.org.br/entenda-o-que-e-assedio-judicial

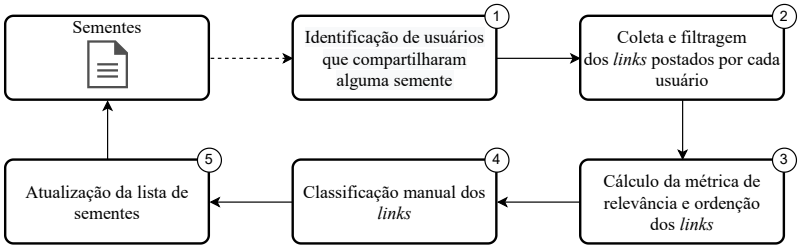

Figura 1. Metodologia para identificação de *websites* de desinformação.

usuários que as tenham compartilhado na plataforma alvo (i.e., usuários que compartilharam uma semente) (Etapa 1). Baseados nesse conjunto de usuários identificados, coletamos todos os *links* também postados pelos mesmos (Etapa 2). É válido destacar que durante esta etapa são removidos *links* que conhecidamente não representam um *website* de notícia (e.g., links para redes sociais ou *websites* governamentais) a fim de minimizar as verificações que serão feitas na próxima etapa. Em seguida, para cada um dos *links* calculamos uma medida de relevância que nos permita capturar a probabilidade de que o mesmo contenha desinformação, e geramos uma lista ordenada, de forma que, *links* mais prováveis sejam apresentados no topo da lista (Etapa 3). Depois, realizamos uma verificação manual dos *websites* ou dos *links* (Etapa 4) com objetivo de identificar novos *links* de notícias para retroalimentação do processo (Etapa 5). Por fim, cada execução completa da metodologia é definida como um *ciclo* que pode ser repetido N vezes.

É importante ressaltar que tanto a lista de *websites* obtida a partir da metodologia proposta neste trabalho quanto as sementes utilizadas não são compartilhadas publicamente, uma vez que o nosso objetivo não é acusar *websites* de produzirem e/ou difundirem desinformação. O objetivo principal deste trabalho é propor uma metodologia que permita que pesquisadores e órgãos competentes possam construir suas próprias listas de *websites* de desinformação (ou baixa credibilidade), sendo estas validadas, por exemplo, a partir de checagens realizadas por agências internacionalmente reconhecidas[3]. Acreditamos que uma lista gerada a partir da aplicação da metodologia proposta ofereça subsídios valiosos para análises de diferentes fenômenos sociais neste contexto.

3. Aplicação da Metodologia Proposta no Twitter

Como uma análise preliminar do potencial da abordagem proposta, aplicamos nossa metodologia para identificação de *websites* de desinformação no Twitter, uma plataforma intensamente utilizada para fomentar discussões sobre diferentes temas relacionados ao cotidiano e amplamente abusada para disseminação de campanhas de disseminação de desinformação [Bovet and Makse 2019]. Apesar dos usuários nessa plataforma interagirem entre si de diversas maneiras (e.g., seguir, retweetar, comentar), o nosso trabalho considera somente as publicações realizadas pelos usuários neste ambiente, uma vez que publicações podem conter *links* para *websites* (de notícias) externos. Após a escolha do total de ciclos, selecionamos a notícia n comprovadamente falsa, cujo *link* será utilizado como semente da execução do primeiro *ciclo*[4] (Etapa 1).

Na Etapa 2, utilizamos a busca da plataforma para coletar o subconjunto de

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

[4]É válido destacar que a primeira execução foi realizada apenas com uma notícia semente, mas que a metodologia proposta permite que várias notícias sejam utilizadas em conjunto durante um mesmo *ciclo*.

usuários U que postaram um *tweet* contendo o *link* para a notícia n. Para cada usuário $u_i \in U$ realizamos uma única coleta da sua *timeline* t_i, contendo as postagens ao longo do tempo, através da API do Twitter[5]. A partir das postagens coletadas, extraímos o conjunto de *links* L_i compartilhados por cada usuário. A cada ciclo, a partir dos conjuntos L_i de links, atualiza-se o conjunto $\mathcal{L}$ de todos os links já coletados. Desse conjunto são retirados os *links* que reconhecidamente não pertencem a um *website* de notícia, bem como aqueles que pertencem a *websites* de notícias de alta credibilidade, ou seja, afiliados a Associação Nacional de Jornais (ANJ) e que seguem um código de ética definido[6].

Depois, na Etapa 3, os links de $\mathcal{L}$ são ordenados em ordem decrescente usando o *h-index* do domínio e o número de usuários que compartilharam o referido *link*, como critério de desempate. Nesse trabalho definimos o *h-index*[7] de cada *website* como o número máximo h, tal que h notícias (*links*) foram publicadas por pelo menos u usuários diferentes (com $u = h$). O cálculo desta métrica é realizado a cada ciclo, dado que novas informações são obtidas, mais *links* são inseridos em $\mathcal{L}$.

Na etapa seguinte (4), realizamos uma classificação manual da lista ordenada até identificarmos um número k de *links* de $\mathcal{L}$ que tenham sido verificados como desinformação (i.e., falsos/enganosos). Nesta etapa[8], definimos $k = 1$. Como estamos buscando *websites* de desinformação, conforme definido anteriormente (Seção 2), consideramos um *link* como de desinformação se existe uma checagem para ele ou para qualquer notícia pertencente ao mesmo *website*. Ao considerar tanto o *link* quanto o *website* ao fazer a verificação, aceleramos o processo de encontrar novos *websites*. No entanto, também é possível que a verificação seja feita considerando somente checagens para a própria notícia (*link*) o que, apesar de tornar o processo de verificação mais custoso, permite que ao final da execução uma lista de notícias e verificações correspondentes seja encontrada. Assim, neste trabalho, utilizamos somente verificações produzidas por algum portal de verificação reconhecido internacionalmente, a saber: Agência Lupa, Aos Fatos e *Estadão Verifica*[9]. Reiteramos que neste trabalho usamos $k = 1$ e que quando um *link* é classificado como desinformação, todos os *links* daquele mesmo *website* são retirados de $\mathcal{L}$ e não são adicionados novamente no futuro.

Por fim, na Etapa 5, adicionam-se os k *links* classificados como desinformação e suas respectivas verificações na lista de sementes. No próximo ciclo, esses k *links* serão usados como sementes. Nesse momento, verifica-se também se c ciclos foram executados. Caso isso seja verdade, a execução da metodologia é finalizada e a lista de sementes (*links* verificados como desinformação) é processada para encontrar os domínios de cada uma, gerando o resultado final dessa execução.

4. Resultados Preliminares

A partir da definição e aplicação da metodologia proposta (Seção 2) na plataforma Twitter (Seção 3), é altamente provável que sejam identificados *websites* de desinformação. No

[5]https://developer.twitter.com/en/docs/twitter-api

[6]https://www.anj.org.br/associados/, https://www.anj.org.br/associe-se/

[7]https://en.wikipedia.org/wiki/H-index

[8]É importante destacar que nesta etapa inicial k foi definido como 1 considerando o custo associado para identificação de uma checagem correspondente. No entanto, a metodologia pode ser estendida para qualquer valor de k.

[9]lupa.uol.com.br, www.aosfatos.org/, politica.estadao.com.br/blogs/estadao-verifica/

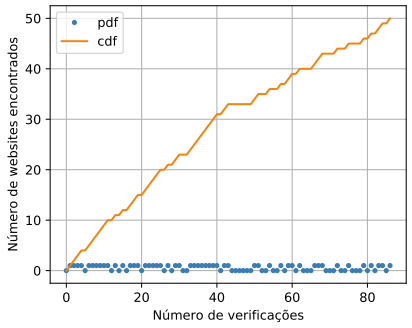

Figura 2. Em laranja, o número de *websites* encontrados em função do número de verificações. Em azul, a função de densidade de verificações.

Número de Verificações	Frequência Absoluta	Frequência Relativa
1	29	0,58
2	13	0,26
3	4	0,08
4	3	0,06
7	1	0,02

Tabela 1. Frequências do número de verificações necessárias até que um *website* de desinformação subsequente seja identificado.

entanto, dentro desse contexto, uma etapa importante e com impacto direto nos resultados é a classificação manual dos *links* – a qual pode ser custosa por envolver trabalho humano. Assim, conduzimos uma análise preliminar com o intuito de estimar o custo (em termos de verificações necessárias) para identificação de um novo *website* de desinformação. Especificamente um custo próximo de 1 indica que foi necessário realizar apenas 1 verificação para identificação de um novo *website* de desinformação, o que seria o cenário ideal.

Primeiramente, partindo de uma notícia semente n, executamos a metodologia para identificação de 50 diferentes *websites* (*ciclos*) que compartilhassem desinformação. Para a descoberta desses *websites* foram necessárias 86 verificações, com um custo médio $1,7$ verificação para cada *website*. A Figura 2 apresenta a relação entre o número de *websites* de desinformação encontrados e o número de verificações, o que nos fornece uma estimativa do custo ao longo dos *ciclos*. De forma geral, podemos observar que após a identificação de 30 novos *websites* há um aumento do custo total acumulado de verificação, de $1,3$ (*ciclos* 1 a 30) para $2,4$ (*ciclos* 31 a 50). Ademais, a Tabela 1 apresenta o total de verificações necessárias até encontrar um *website* de desinformação. Idealmente, espera-se que esse número seja baixo, isto é, o mais próximo possível de 1. Podemos observar que em 84% dos *ciclos* foi necessário executar no máximo 2 verificações e em 58% dos *ciclos* apenas uma validação foi suficiente. Com os dados obtidos, notamos um custo pequeno no processo de verificação. No entanto, acreditamos que mais análises são necessárias para entendimento aprofundado do comportamento do custo em função dos *ciclos* executados, e assim considerar definir um numero máximo de verificações por *ciclo* ou de *ciclos* por execução. Por fim, uma comparação com abordagens alternativas neste contexto, é uma direção que também pretendemos explorar em trabalhos futuros.

5. Conclusão

Nesse trabalho, apresentamos uma metodologia para a identificação de *websites* de desinformação (i.e., baixa credibilidade) no Brasil. A principal contribuição do nosso trabalho está em criar uma metodologia e não em prover uma lista de *websites* identificados. O motivo de tal lista não ser disponibilizada no trabalho é que um *website* identificado na lista poderia nos processar como forma de intimidação. Já a metodologia apresentada é imune ao assédio judicial, tendo em vista que ela não acusa nenhum

website específico, mas permite que pesquisadores e entidades governamentais possam construir suas próprias listas de *websites*. Nesse sentido, acreditamos que a metodologia aqui proposta abre uma nova avenida de possibilidades, conforme discutimos a seguir.

Desinformação no Brasil. Acreditamos que uma série de trabalhos relacionados à desinformação possam surgir no Brasil a partir da nossa metodologia. Como exemplos de direções, podemos procurar o melhor entendimento e espalhamento desses *websites* nas diversas plataformas digitais que incluem não só o Twitter, mas também aplicativos para troca de mensagens, como WhatsApp e Telegram. Além disso, há uma série de aspectos desses *websites* que podem ser explorados que vão desde os seus registros no PontoBR até aspectos relacionados à respectiva monetização.

Atuação do Poder Público. Esperamos que nossa metodologia possa ser utilizada por agentes do poder público que podem investigar quem cria, financia e se beneficia pela existência de *websites* dedicados à produção e disseminação de desinformação em plataformas digitais. Atualmente, nosso grupo está trabalhando em cooperação com o Ministério Público de Minas Gerais (MPMG) no sentido de oferecer material que justifique a abertura de investigação de tais *websites*.

Agradecimentos. Este trabalho foi parcialmente financiado pelo MPMG, projeto Capacidades Analíticas, CNPQ, FAPEMIG e FAPESP.

Referências

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.

Bessi, A. and Ferrara, E. (2016). Social bots distort the 2016 us presidential election online discussion. *First monday*, 21(11-7).

Bovet, A. and Makse, H. A. (2019). Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):7.

Bozarth, L. and Budak, C. (2021). Market forces: Quantifying the role of top credible ad servers in the fake news ecosystem. In *Proc. of the Int'l AAAI Conf. on Web. and Soc. Med. (ICWSM)*, pages 83–94.

Galhardi, C. P., Freire, N. P., Minayo, M. C. d. S., and Fagundes, 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.

Martins, A. D. F., Monteiro, J. M., and Machado, J. (2021). Automatic misinformation detection about covid-19 in brazilian portuguese whatsapp messages. In *Proc. of the Brazilian Symposium on Data Bases (SBBD)*, pages 120–126.

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L., Recchia, G., Van Der Bles, A. M., and Van Der Linden, S. (2020). Susceptibility to misinformation about covid-19 around the world. *Royal Society open science*, 7(10):201199.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., and Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42.