Capítulo 1

Desinformação em Plataformas Digitais: Conceitos, Abordagens Tecnológicas e Desafios

Julio C. S. Reis, Philipe Melo, Márcio Silva, Fabrício Benevenuto

Abstract

Digital platforms, including online social networks and instant messaging applications, have become spaces widely abused by misinformation campaigns, directly impacting various spheres of our society such as politics, health, and others. Consequently, this has stimulated the emergence of research on this topic in several areas of knowledge, including Computer Science. Thus, the general goal of this chapter is to discuss the current scenario of studies in the context of misinformation on digital platforms, offering an introduction to the researcher to explore this theme. Thus, we first present and discuss the concepts underlying the area. Then, we list data repositories that may be useful for studying the phenomenon. Afterward, we describe the main strategies explored for understanding as well as technological approaches for detecting and monitoring disinformation on digital platforms. Last, we present a critical overview of the area, highlighting research challenges and opportunities in this context.

Resumo

Plataformas digitais, que incluem redes sociais online e aplicativos de mensagem instantânea, se tornaram espaços amplamente abusados por campanhas de desinformação com impactos diretos em diversas esferas da nossa sociedade, incluindo política, saúde, dentre outras. Consequentemente, isso tem estimulado o surgimento de pesquisas neste tema em diversas áreas do conhecimento, incluindo Ciência da Computação. Assim, o objetivo geral deste capítulo é discutir o cenário atual de estudos no contexto de desinformação em plataformas digitais, oferecendo uma introdução ao pesquisador que pretende explorar este tema. Para isso, inicialmente, são apresentados e discutidos os conceitos que fundamentam a área. Em seguida, são relacionados repositórios de dados que podem ser úteis para o estudo deste fenômeno. Depois, são sumarizadas as principais estratégias exploradas para entendimento, bem como abordagens tecnológicas para detecção e monitoramento de desinformação em plataformas digitais. Por fim, é apresentada uma visão crítica geral da área, destacando desafios e oportunidades de pesquisa neste contexto.

1.1. Introdução

As plataformas digitais, que incluem redes sociais online como Facebook, Instagram e Twitter, e aplicativos para troca de mensagens instantâneas como WhatsApp e Telegram, são usadas ativamente por mais da metade da população mundial [Gallagher, 2017]. Essas plataformas mudaram significativamente a forma como as pessoas interagem e se comunicam no ambiente online modificando vários dos ecossistemas de informação existentes. Particularmente, as plataformas digitais mudaram drasticamente a maneira como as notícias são produzidas, disseminadas e consumidas em nossa sociedade, abrindo oportunidades imprevistas e também criando desafios complexos.

Parte das razões para esta mudança é inerente à natureza das próprias plataformas digitais: *(i)* muitas vezes é mais oportuno e menos dispendioso produzir e consumir notícias nesses ambientes em comparação com meios noticiosos tradicionais, como jornais online e/ou físicos, ou ainda rádio e televisão; e *(ii)* é mais fácil compartilhar, comentar e discutir as notícias com amigos ou outros leitores em plataformas digitais, o que melhora e/ou impulsiona a comunicação e as interações entre os usuários nesses sistemas [Shu et al., 2017]. Assim, as plataformas digitais estão moldando a maneira como as pessoas consomem informações, especialmente notícias. Um estudo revelou que cerca de 62% dos usuários americanos e 66% dos usuários brasileiros recebem e consomem notícias a partir desses sistemas [Mitchell, 2016; Report, 2018]. Apesar dos inúmeros benefícios que essas plataformas trazem para nossa sociedade, elas se tornaram um local propício para realização de campanhas de desinformação que muitas vezes visam enganar as pessoas, especialmente em contextos como saúde e política.

Em relação a saúde, notícias "médicas" contendo desinformação divulgadas em plataformas digitais causaram danos irreparáveis [Dai et al., 2020]. Por exemplo, na China, um paciente com câncer confundiu um anúncio online com um tratamento experimental contra a doença, acreditando ser uma informação clinicamente confiável, o que infelizmente resultou em sua morte [Dai et al., 2020][1]. Além disso, durante a pandemia de COVID-19, houve um aumento significativo dos rumores e conspirações espalhados pelas plataformas digitais [Ferrara, 2020]. A *International Fact-Checking Network* (IFNC)[2] encontrou mais de 3.500 alegações falsas relacionadas a COVID-19 em menos de dois meses [Poynter, 2020]. Como resultado, pelo menos 800 pessoas podem ter morrido em todo o mundo por causa de desinformação relacionada ao coronavírus nos primeiros três meses de 2020[3].

Já no contexto político, eleição após eleição, podemos observar diferentes formas de desvio de conduta e estratégias complexas de manipulação de opinião por meio da disseminação de desinformação em plataformas digitais. A eleição presidencial de 2016 nos EUA ainda é lembrada pela "guerra de desinformação" que aconteceu principalmente por meio do Twitter e do Facebook. Um caso notório envolveu uma tentativa de influência da Rússia por meio de publicidade segmentada [Ribeiro et al., 2019]. Tentativas semelhantes foram observadas durante as eleições brasileiras de 2018, onde o WhatsApp foi extensivamente abusado para envio de campanhas de desinformação, com grande uso de

[1] https://www.bbc.com/news/business-36189252
[2] https://www.poynter.org/ifcn/
[3] https://www.bbc.com/news/world-53755067

imagens e memes manipulados[4] contendo todo tipo de ataque político. Por exemplo, um estudo mostrou que 88% das imagens mais populares compartilhadas no último mês antes das eleições brasileiras de 2018 eram falsas ou enganosas [Tardaguila et al., 2018]. Também por meio do WhatsApp, na Índia, boatos falsos espalhados pela plataforma foram responsáveis por vários casos de linchamento e agitação social [Arun, 2019].

Uma característica única das notícias em plataformas digitais que suportam esse fenômeno da disseminação da desinformação nesses ambientes é que qualquer pessoa pode se registrar e/ou comportar-se como um editor de notícias sem nenhum custo inicial (e.g., qualquer pessoa pode criar uma página no Facebook alegando ser um jornal ou organização de mídia de notícias, ou ainda, criar um grupo no WhatsApp ou Telegram para divulgação deste tipo de conteúdo). Consequentemente, não apenas as empresas de notícias tradicionais (e.g., jornais) estão migrando para plataformas digitais, mas também muitos veículos de notícias também estão emergindo exclusivamente nesses ambientes[5]. Por exemplo, esforços anteriores mostraram que em 2018 havia mais de 20 mil páginas nos EUA categorizadas como editores de notícias no Facebook [Ribeiro et al., 2018], e este número certamente continua crescendo.

Junto à esta transição, há uma preocupação crescente sobre como os produtores de notícias contendo desinformação elaboram e publicam este tipo de conteúdo[6], muitas vezes divulgando-as amplamente através de plataformas digitais [Lazer et al., 2018]. Por exemplo, um estudo financiado pela Avaaz[7] perguntou aos eleitores brasileiros se eles viram e acreditaram em cinco das notícias mais populares contendo desinformação e disseminadas em plataformas digitais durante as últimas semanas da eleição presidencial em 2018. De forma impressionante, os resultados revelaram que mais de 98% dos eleitores entrevistados foram expostos a uma ou mais dessas notícias contendo desinformação e que quase 90% deles acreditaram que essas histórias eram verdadeiras[8]. Potencialmente, esses números impactaram a democracia no Brasil diante das eleições presidenciais de 2018.

Desinformação, manipulação de opinião, mentiras, boatos, rumores, e enganos sempre existiram, mas a ascensão das plataformas digitais aumentou significativamente o potencial da disseminação deste tipo de conteúdo transformando este problema em um fenômeno mundial, que tem atraído a atenção de pesquisadores de diversas áreas, incluindo Ciência da Computação. Isso impulsiona a necessidade de discussões sobre o impacto das plataformas digitais frente a este fenômeno cada vez mais desafiador.

Assim, acreditamos que esse é essencial fornecer às pessoas insumos para: *(i)* suporte ao entendimento do fenômeno da desinformação considerando diferentes contextos, cenários, e ambientes; *(ii)* proposição de abordagens automatizadas que possam ser

[4]"Uma imagem, vídeo, texto, etc., tipicamente de natureza humorística, que é copiado e difundido rapidamente pelos internautas, muitas vezes com pequenas variações" [Oxford, 2020].

[5]https://www.comscore.com/Insights/Blog/Traditional-News-Publishers-Take-Non-Traditional-Path-to-Digital-Growth

[6]https://www1.folha.uol.com.br/ilustrissima/2017/02/1859808-como-funciona-a-engrenagem-das-noticias-falsas-no-brasil.shtml

[7]https://www.avaaz.org/

[8]https://www1.folha.uol.com.br/poder/2018/11/90-dos-eleitores-de-bolsonaro-acreditaram-em-fake-news-diz-estudo.shtml

efetivas na contenção e/ou detecção deste tipo de conteúdo, e por fim; *(iii)* embasamento ao processo de tomada de decisão por parte dos indivíduos até entidades e órgãos de gestão das plataformas digitais e governamentais. De forma geral, esperamos contribuir com esforços focados na minimização dos impactos ocasionados pelo problema em nossa sociedade.

Neste sentido, este capítulo visa não só uma exposição teórica dos conceitos e definições relacionadas à temática, mas também apresentar práticas introdutórias, as quais podem ser consideradas passos fundamentais no estudo e desenvolvimento de soluções no contexto da desinformação em plataformas digitais, contribuindo com a formação de recursos humanos, capacitação e compartilhamento de conhecimentos a respeito dos impactos relacionados ao fenômeno. Além disso, esperamos fornecer aos leitores uma compreensão abrangente dos conceitos relacionados à disseminação da desinformação em plataformas digitais, como redes sociais e aplicativos de mensagens.

Este capítulo está organizado da seguinte forma. Primeiramente, apresentamos brevemente o estado da arte nesta área, incluindo definições e uma descrição do ecossistema de (des)informação nessas plataformas, bem como áreas correlatas. Em seguida, relacionamos os principais repositórios de dados e estratégias atualmente explorados para entendimento e caracterização do problema, bem como abordagens tecnológicas para mitigar seus potenciais impactos na sociedade. Além disso, este capítulo também inclui um relato de experiência de um projeto real, o ELEIÇÕES SEM FAKE, desenvolvido em parceria com o Tribunal Superior Eleitoral (TSE) do Brasil e amplamente utilizado para enfrentamento à desinformação durante o processo eleitoral no país. Por fim, apresentamos considerações finais, incluindo desafios e oportunidades de pesquisa na área.

1.2. Ecossistema de Notícias em Plataformas Digitais

A mídia noticiosa tem sido objeto de estudo de diversas áreas do conhecimento, como Jornalismo, Comunicação e Ciências Políticas. No entanto, desde que ela se concentrou na Web e nas plataformas digitais, este tem sido também um tópico de interesse dos cientistas da computação. Com o surgimento da era digital, os meios de comunicação começaram a publicar no ambiente online. Assim, com o farto rastro digital de informações jornalísticas, as possibilidades de novas aplicações e o surgimento de novos desafios nesse cenário complexo, cientistas da computação têm investigado problemas relacionados ao ecossistema de notícias em plataformas digitais, mas geralmente com objetivos e finalidades diferentes.

Em suma, a base do ecossistema de notícias nas plataformas digitais pode ser dividida em três componentes principais: *(i) produção*, *(ii) consumo* e *(iii) disseminação e interação*, conforme apresentado na Figura 1.1. Primeiro, antes das plataformas digitais, os artigos de notícias eram produzidos (ou escritos) apenas por organizações tradicionais de mídia (i.e., jornais) ou por jornalistas independentes. Com o surgimento das plataformas digitais, uma das principais características da *(i) produção* de notícias nesses ambientes é que qualquer um pode ser um produtor de notícias. Por exemplo, qualquer um pode criar um usuário em uma plataforma digital para produzir e divulgar notícias sem nenhum custo inicial. Ademais, o *(ii) consumo* de notícias também mudou ao longo do tempo de papel de jornal (físico) para rádio/televisão e, depois, para notícias online

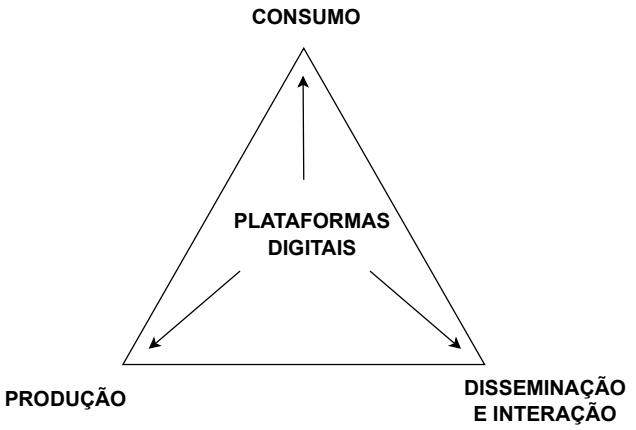


Figura 1.1. Ecossistema da informação em plataformas digitais.

e plataformas digitais, onde muitas vezes é mais oportuno e barato consumir notícias em comparação com a mídia tradicional. Por exemplo, uma pesquisa do *Pew Research Center* estima que 62% dos adultos nos EUA consomem notícias principalmente de sites de mídia social [Mitchell, 2016]. No Brasil, segundo pesquisa realizada pelo *Reuters Institute*, esse percentual chega a 66% [Report, 2018]. Por fim, as plataformas digitais introduzem novos mecanismos para *(iii) disseminação e interação*, permitindo que os usuários compartilhem e promovam notícias de acordo com sua vontade, o que pode ser comitantemente benéfico e perigoso.

Logo, em decorrência da inserção das plataformas digitais no ecossistema de notícias, existem diferentes esforços no contexto da Ciência da Computação que buscam entender melhor essas mudanças e propor soluções para dar suporte a este fenômeno em suas diversas fases. Esses esforços podem ser agrupados em três conjuntos relacionados aos principais componentes do ecossistema de notícias em plataformas digitais. O primeiro aborda a *(i) produção* de conteúdo e envolve temas como cobertura e eventos de notícias [Kwak and An, 2014; Quezada et al., 2015], credibilidade [Jin et al., 2014], aspectos da atratividade [Kim et al., 2016; Reis et al., 2015], e viés das notícias [Covert and Wasburn, 2007; Budak et al., 2016]. O segundo conjunto está relacionado ao *(ii) consumo* e envolve estudos relacionados ao entendimento de padrões de leitura das pessoas [Kourogi et al., 2015; Constantinides et al., 2015], personalização de conteúdo [Das et al., 2007; Garcin et al., 2014], sumarização e/ou resumo de notícias [Gao et al., 2012], sistemas de recomendação [Nallapati et al., 2004], visualização de conteúdo [Sheidin et al., 2017] e consumo de notícias no celular e dispositivos [Westlund, 2013; Constantinides, 2015]. Por fim, o terceiro está associado aos mecanismos de *(iii) disseminação e interação* proporcionados pelas plataformas digitais, incluindo esforços dedicadas à compreensão desses mecanismos [Chakraborty et al., 2016], motivações para os usuários compartilharem [Lee and Ma, 2012] e novos desafios que emergem desses mecanismos como bolhas de filtro e efeito de câmaras eco [Bakshy et al., 2015].

1.3. Visão Geral de Desinformação

Conforme supracitado, a mídia noticiosa adentrou as plataformas digitais e tem sido um tópico de interesse de vários pesquisadores, incluindo os cientistas da computação. No entanto, as mudanças no ecossistema da mídia jornalística ainda acontecem rapidamente, e

(a) Versão 1

(b) Versão 2

Figura 1.2. Exemplo de desinformação abordando um tratamento alternativo, milagroso e fácil para a cura do câncer. Fonte: `https://observatoriofakenews.eci.ufmg.br/2018/08/01/limonada-quente-destroi-celulas-cancerigenas/`.

algumas delas favorecem campanhas de desinformação, revelando as plataformas digitais como ambientes potenciais e/ou propícios para a disseminação deste tipo de conteúdo.

A Figura 1.2 mostra um exemplo de diferentes versões de uma mesma notícia falsa bastante popular (i.e., contendo desinformação) e propagada em plataformas digitais como Facebook, Twitter e WhatsApp sobre o potencial do suco de limão quente para a cura do câncer. A alegação foi verificada como "falsa" por várias agências de verificação de fatos em todo o mundo, incluindo Snopes[9], e Boatos.org[10] no Brasil. Especificamente sobre essa notícia contendo desinformação, a agência de verificação de fatos Snopes concluiu que "o melhor que pode ser dito é que as frutas cítricas podem potencialmente abrigar propriedades anticancerígenas que podem ajudar a prevenir o câncer. Nenhum estudo científico ou médico respeitável relatou que os limões foram definitivamente considerados um 'remédio comprovado contra cânceres de todos os tipos', nem nenhum dos (convenientemente sem nome) 'maiores fabricantes de medicamentos do mundo' relatou a descoberta de que os limões são '10.000 vezes mais forte que a quimioterapia' e que sua

[9] `https://www.snopes.com/fact-check/lemon-cancer-cure/`
[10] `https://www.boatos.org/saude/limonada-quente-mata-cancer.html`

ingestão pode 'destruir células malignas [câncer]'. Todas essas alegações são hipérboles e exageros não suportados por fatos"[11].

Esforços anteriores sugerem que existem pelo menos três tipos de notícias falsas ou desinformação [Rubin et al., 2015]. O primeiro tipo consiste em *(i)* sátira ou paródia, onde *websites* como o Onion[12] ou Daily Mash[13] publicam notícias muitas vezes contendo desinformação como tentativas humorísticas de satirizar a mídia. O segundo tipo *(ii)* contempla notícias que apresentam desinformação (i.e., informações falsas) em conjunto com informações (parcialmente) verdadeiras, mas usadas em contexto errado, incluindo boatos e notícias enganosas que não são baseadas em fatos, mas sustentam uma narrativa contínua. Por último, o terceiro grupo envolve *(iii)* notícias criadas intencionalmente com desinformação. Normalmente, elas são fabricadas e divulgadas deliberadamente em plataformas digitais para obtenção de lucros financeiros a partir, por exemplo, do número de cliques, ou ainda, para enganar, causar confusão e/ou manipular a opinião pública[14][15]. Neste capítulo, nos aprofundamos no grupo *(iii)*.

1.3.1. Definição de Desinformação

Desinformação é um tema que ainda carece de uma definição clara ou universalmente aceita. De acordo com o dicionário Collins English[16] o termo "notícias falsas" (do inglês "*fake news*") por ser definido, por exemplo, como "*informações falsas, muitas vezes sensacionalistas, disseminadas sob o disfarce de notícias*". No entanto, a definição destes termos (i.e., "desinformação", "notícias falsas", "*fake news*", etc), bem como a sua percepção, conceptualização e relação, tem sido objeto de debate recente [Shu et al., 2017]. Logo, é crucial estabelecermos a definição para o termo que será utilizada ao longo deste capítulo.

Definição 1.3.1 ***(Desinformação)*** *"Uma notícia ou mensagem publicada e propagada pela mídia, contendo informações falsas, independentemente dos meios e motivos que a embasa" [Sharma et al., 2019].*

1.3.2. Desinformação em Plataformas Digitais

De forma geral, conforme mencionado anteriormente, o ecossistema de notícias, que abrange as contendo desinformação, foi mudando ao longo do tempo do jornal impresso para o rádio/televisão e, depois, para notícias online e mais recentemente, plataformas digitais. No entanto, existem vários fundamentos sociais e teorias psicológicas e cognitivas que descrevem o impacto da desinformação tanto no nível individual quanto no ecossistema de informações sociais. Primeiro, os leitores preferem receber informações que confirmem suas opiniões existentes [Nickerson, 1998]. Em segundo lugar, os usuários fazem escolhas com base nos ganhos e perdas relativos em comparação com seu estado atual [Tversky and Kahneman, 1992]. Finalmente, os leitores tendem a acreditar que suas

[11] Tradução livre dos autores.

[12] `www.theonion.com`

[13] `www.thedailymash.co.uk`

[14] `http://www.cits.ucsb.edu/fake-news/danger-election`

[15] `https://www1.folha.uol.com.br/ilustrissima/2017/02/1859808-como-funciona-a-engrenagem-das-noticias-falsas-no-brasil.shtml`

[16] `https://www.collinsdictionary.com/woty`

percepções da realidade são as únicas visões precisas, enquanto outros que discordam são considerados desinformados, irracionais ou tendenciosos/enviesados [Ward et al., 1997]. Todos esses fatores potencializam a disseminação de desinformação pelos usuários no contexto das plataformas digitais. Além disso, existem outras características relacionadas às estes sistemas que contribuem para a disseminação da desinformação nesses ambientes, as quais serão discutidas a seguir.

Contas Maliciosas. Os usuários em plataformas digitais podem ser legítimos ou não. O baixo custo de criação de contas em plataformas digitais encorajou contas de usuários maliciosos [Shu et al., 2017], como *bots* sociais e *trolls*, que são controlados por um algoritmo de computador para interagir automaticamente com humanos (ou outros *bots*) nestes sistemas [Ferrara et al., 2016]. Nesse contexto, muitos *bots* são criados especificamente com a finalidade de causar danos, incluindo a manipulação e o espalhamento de desinformação em plataformas digitais. Existem esforços anteriores que discutem como os *bots* sociais impactaram, por exemplo, a discussão online relacionada às eleições presidenciais americanas de 2016 [Bessi and Ferrara, 2016; Badawy et al., 2019], ou ainda, como eles coordenaram campanhas de desinformação durante as eleições presidenciais francesas de 2017 [Ferrara, 2017]. Outros estudos mostram que os *bots* foram responsáveis por aumentar significativamente a disseminação de desinformação nas plataformas digitais, sugerindo que coibir os *bots* sociais pode ser uma estratégia eficaz para conter o problema neste ambiente [Shao et al., 2018; Wang et al., 2018a].

Publicidade em Mídias Digitais. Nos últimos anos, as plataformas de publicidade em mídias digitais evoluíram significativamente [Silva et al., 2020]. Com acesso a informações pessoais e atividades de milhões de pessoas ao redor do mundo, esses ambientes permitem que os anunciantes atinjam nichos muito específicos de usuários considerando informações pessoais como nome, endereço de e-mail, aspectos demográficos, comportamentos e muitos outros. No entanto, a publicidade direcionada nesses sistemas também pode ser utilizada de forma abusiva por anunciantes mal-intencionados para alcançar com eficiência pessoas suscetíveis e/ou vulneráveis a histórias falsas, alimentar queixas e incitar conflitos sociais [Ribeiro et al., 2019]. Esforços anteriores destacaram várias formas de abuso de publicidade direcionada em plataformas como Facebook, por expor informações privadas dos usuários de forma inadequada à anunciantes e por permitir publicidade discriminatória (e.g., excluir pessoas de uma determinada raça ou gênero de receber seus anúncios [Andreou et al., 2018; Venkatadri et al., 2018]). Além disso, outros estudos investigaram até que ponto os anúncios políticos da Agência de Pesquisa de Inteligência Russa (IRA) veiculados antes das eleições de 2016 nos EUA exploraram a infraestrutura de publicidade direcionada do Facebook para direcionar anúncios com eficiência em tópicos divisivos ou polarizadores (e.g., imigração, raça, etc) em subpopulações vulneráveis [Ribeiro et al., 2019]. De forma geral, os resultados sugerem que plataformas de anúncio e/ou propagandas estão sendo exploradas para uma nova forma de ataque, que é o uso de publicidade direcionada para criar discórdia social e atingir pessoas suscetíveis a informações específicas, incluindo histórias contendo desinformação.

Efeito Câmara de Eco. Finalmente, como mencionado anteriormente, as plataformas digitais deram origem a novos fenômenos disruptivos no ecossistema de notícias: as chama-

das câmaras de eco[17]. As câmaras de eco referem-se a grupos dentro de uma plataforma digital onde as pessoas (ou leitores) raramente são expostos a conteúdos que atravessam linhas ideológicas, mas são alimentados com informações que reforçam suas visões políticas ou sociais atuais pré-existentes. Embora em alguns casos a classificação algorítmica (que decide o que é apresentado no *feed* – *timeline* – ou nos resultados de pesquisa de alguém e em qual ordem) possa contribuir para esse efeito, pesquisas baseadas em dados do Facebook mostraram que as escolhas dos indivíduos são o principal fator para limitar a exposição à conteúdo transversal [Bakshy et al., 2015]. Pesquisas anteriores têm mostrado que o efeito câmara de eco facilita o acesso pelo qual as pessoas consomem e acreditam em notícias contendo desinformação devido, inclusive, a fatores como: *(i)* os relacionamentos dos usuários influenciam na confiabilidade deles em determinada fonte da informação, ou seja, os usuários tendem a perceber uma fonte como confiável se outros também perceberem-na como confiável [Shu et al., 2017]; *(ii)* os leitores podem naturalmente favorecer informações que ouvem com frequência, mesmo que sejam notícias contendo desinformação [Zajonc, 1968], e; *(iii)* nas câmaras de eco, os usuários continuam compartilhando e consumindo as mesmas informações [Shu et al., 2017]. Como resultado, esse efeito de câmara de eco cria comunidades segmentadas, homogêneas e ideologicamente polarizadas com um ecossistema de informações muito limitado, o que consequentemente favorece campanhas de desinformação [Sasahara et al., 2020].

1.3.2.1. Áreas Correlatas

Por fim, existem algumas outras áreas relacionadas ao tema de desinformação, com algumas especificidades e/ou eventuais sobreposições. Como exemplos, podemos destacar: a *(i) identificação de rumores*, que permeia história ou declaração de circulação geral, sem confirmação ou certeza dos fatos [Allport and Postman, 1947]; a *(ii) descoberta de verdade*, que visa determinar a credibilidade da fonte e a veracidade do objeto ao mesmo tempo [Li et al., 2016]; *(iii) Clickbait* que se refere a "conteúdo cujo objetivo principal é atrair a atenção e encorajar os visitantes a clicar em um link para uma determinada página da Web"[18]; ou ainda a chamada era da *(iv) pós-verdade*, em que fatos objetivos têm menos relevância na formação da opinião pública do que os apelos individuais[19], dentre outros. Logo, estas áreas e outros aspectos correlatos podem ser explorados comutantemente abrangendo, inclusive, medidas que podem ser tomadas para combater os impactos ocasionados pela disseminação da desinformação no contexto das plataformas digitais.

1.4. Repositório de Dados

Para que sejamos capazes de realizar contribuições concretas para a compreensão e detecção de desinformação em plataformas digitais, precisamos de amplos repositórios de dados contendo instâncias rotuladas por especialistas, ou seja, fatos e conteúdo verificado, abrangendo diferentes tópicos e contextos [Hui et al., 2018; Nørregaard et al., 2019].

Diante disso, realizamos um breve levantamento sobre conjuntos de dados públicos existentes comumente utilizados por trabalhos que investigam o fenômeno da desin-

[17] https://cs181journalism2015.weebly.com/the-echo-chamber-effect.html
[18] http://www.oxforddictionaries.com/definition/english/clickbait
[19] https://languages.oup.com/word-of-the-year/2016/

formação em plataformas digitais, seja para entendê-lo ou para propor soluções que visem minimizar os efeitos por ele causados. Esses conjuntos de dados rotulam o conteúdo geralmente como histórias contendo desinformação (*fake news*) ou verdadeiras. Esse conteúdo verificado pode aparecer em vários formatos diferentes, como artigos de notícias, declarações ou citações de celebridades, rumores, relatórios ou imagens e para diferentes cenários, como eleições, saúde, guerras, política. A Tabela 1.1 resume alguns dos principais conjuntos de dados públicos e suas características, incluindo uma descrição, o número total de instâncias, bem como sua distribuição por rótulo (ou seja, veredito fornecido por uma agência de checagem de fatos – verdadeiro; falso; etc.[20]) e informações sobre avaliadores (ou seja, rotuladores e/ou verificadores de fatos). Observe que colorimos em vermelho o número de instâncias verificadas e rotuladas como "desinformação" (ou falsa), em azul, as informações (ou notícias) verdadeiras, e em preto as restantes (por exemplo aquelas que são neutras).

Em alto nível, esses conjuntos de dados de conteúdo checado foram rotulados de acordo com diferentes escalas, como falso ou verdadeiro por jornalistas especializados, sites de verificação de fatos e detectores da indústria[30] fornecendo diferentes informações e contextos que nos permitem extrair características distintas [Shu et al., 2017]. Também podemos observar que, a maioria dos conjuntos de dados relacionados se concentra em notícias políticas dos EUA, notícias de entretenimento ou artigos de sátira com informações extras de sistemas tradicionais como Twitter e Facebook. Logo, ainda há uma carência na literatura de conjuntos de dados que cubram diferentes cenários de interesse, como por exemplo, eleições brasileiras, e abranjam diferentes plataformas, como aplicativos de mensagem instantânea. Isso tem motivado o nosso grupo de pesquisa no desenvolvimento de coletores de dados do WhatsApp e Telegram, conforme apresentado a seguir.

1.4.1. Arquitetura de Coleta de Dados para Aplicativos de Mensagens Instantâneas

Aplicativos de mensagens instantâneas, como o WhatsApp e o Telegram, possuem uma estrutura muito própria, bem diferente de outras plataformas. Enquanto sistemas mais tradicionais como Twitter e Facebook possuem uma esfera pública de conteúdo, onde qualquer um pode acessar e visualizar o que um grande número de usuários está fazendo na rede, suas publicações ou o perfil dos usuários, no WhatsApp e Telegram, as interações e conversas ocorrem um ambiente muito mais restrito. Neste cenário, as mensagens são descentralizadas em distintos grupos e chats de conversa e dificilmente uma pessoa tem acesso a uma parcela significativa do que os outros usuários fazem na rede.

Por isso, a arquitetura da coleta para essas plataformas possui características bem próprias que vão além de APIs, sendo necessário desde um celular com uma conta ativa na rede, até uma estratégia de localização de grupos públicos de conversa para participar, e métodos para extração, processamento e armazenamento dos dados da plataforma. Tudo isso é necessário para identificar e salvar as mensagens que circulam dentro destes ambientes. Portanto, antes de descrevermos detalhes relativos à implementação e instalação de ferramentas para auxiliar o processo de extração e processamento de dados

[20]Neste contexto é importante mencionar que é bastante vasta a nomenclatura utilizada por cada uma das agências de checagem de fatos para indicação do veredito. Logo, optamos por manter a nomenclatura original – sem tradução.

[30]BS Detector: `http://bsdetector.tech/`.

Tabela 1.1. Visão geral dos repositórios de dados rotulados disponíveis na literatura.

Repositório de Dados	Descrição	Rótulos	Rotuladores	# Instâncias
BuzzFace[21] [Potthast et al., 2018; Santia and Williams, 2018]	Notícias publicadas no Facebook por 9 agências ao longo de uma semana perto da eleição de 2016 nos EUA.	mostly false (104), mixture of true and false (245), mostly true (1669), no factual (264)	Jornalistas especialistas do BuzzFeed.	2.282
Central de Fatos [Couto et al., 2021; Marques et al., 2022]	Checagens de fatos sobre assuntos diversos de 6 agências brasileiras.	falso/fake, verdadeiro, enganoso, fato, etc	Agências de checagem de fatos.	11.647
CoAID[22] [Cui and Lee, 2020]	Notícias e reivindicações relacionadas ao COVID-19 em *websites* e plataformas sociais, juntamente com o engajamento social dos usuários sobre essas notícias.	fake (162), true (1.734)	Meios de comunicação confiáveis e sites de verificação de fatos.	1.896
Fact-Checked Images-WhatsApp [Reis et al., 2020b]	Imagens disseminadas no WhatsApp durante períodos eleitorais no Brasil e na India.	misinformation (1.032), not misinformation (780), random (1.261)	Agências de checagem de fatos.	3.073
Fact-Checked-Stat [Vlachos and Riedel, 2014]	Declarações verificadas de sites populares de verificação de fatos rotulados por jornalistas.	true (32), mostly true (34), half true (68), mostly false (37), false (49), fiction (1)	Jornalistas de agências de checagem de fatos.	221
FakeHealth [Dai et al., 2020]	Um repositório que consiste em dois conjuntos de dados, ou seja, Health-Story e HealthRelease e inclui conteúdos de notícias sobre saúde, análises de notícias, engajamentos sociais e redes de usuários.	fake (763), true (1.533)	Revisores especialistas no domínio da saúde.	2.296
Fake.Br Corpus [Monteiro et al., 2018]	Notícias verdadeiras e falsas que foram alinhadas manualmente, focando apenas no português brasileiro.	fake (3,600), true (3.600)	Pesquisadores.	7.200
Fake-News-Net[23] [Shu et al., 2017]	Um repositório para um projeto de coleta de dados em andamento para pesquisa de notícias falsas, incluindo conteúdo de notícias e recursos de contexto social com rótulos de notícias falsas de verdade de grupo confiáveis.	fake (211), real (211)	Jornalistas especialistas do BuzzFeed e checadores de fato do PolitiFact.com.	422
Fake-Real-News[24]	Artigos de notícias publicados durante 2015-2016, juntamente com seus títulos. Todo o corpus é construído a partir de notícias reais coletadas do *The New York Times*[25] e NPR[26], e notícias falsas de um conjunto de dados do Kaggle para garantir uma distribuição uniforme das amostras de ambas as classes.	fake (3.164), real (3.171)	Jornalistas para as notícias verdadeiras e anotadores humanos do BS Detector para notícias falsas.	6.335
Fake-Satire [Golbeck et al., 2018]	Conjunto de dados de notícias falsas e sátiras que são codificadas manualmente, verificadas e, no caso de notícias falsas, incluem histórias refutadas.	fake news (283), satire (203)	Pesquisadores com base em um artigo de um site de verificação de fatos ou em uma informação que refuta uma afirmação.	486
FA-KES [Salem et al., 2019]	Um conjunto de dados de notícias falsas sobre a guerra na Síria (ou seja, relatórios sobre incidentes de guerra ocorridos de 2011 a 2018).	fake (378), true (426)	Abordagem de rotulagem de verificação de fatos semi-supervisionada.	804
Fake-Twitter-Science [Vosoughi et al., 2018]	Todas as notícias verdadeiras e falsas verificadas distribuídas no Twitter de 2006 a 2017. Os dados compreendem ∼126.000 ocorrências (rumores em cascata) tuitadas por ∼3 milhões de pessoas mais de 4,5 milhões de vezes.	true (24.409), false (82.605), mixed (19.287)	Consenso entre verificadores de fatos de seis organizações independentes de checagem de fatos.	126.301
Kaggle[27]	Texto e metadados de fontes de notícias falsas e tendenciosas na Web de BS Detector[28].	bias (443), bs (11.492), conspiracy (430), fake (19), hate (246), junksci (102), satire (146), state (121)	Anotadores humanos do BS Detector.	12.997
LIAR [Wang, 2017]	Declarações curtas do PolitiFact.com rotuladas manualmente.	half-true (2638), false (2511), mostly-true (2466), barely-true (2108), true (2063), pants-fire (1050)	Checadores de fatos do PolitiFact.com.	12.836
NELA-GT-*[29] [Nørregaard et al., 2019; Gruppi et al., 2020]	Artigos de notícias de vários meios de comunicação, incluindo fontes convencionais, hiperpartidárias e de conspiração.	unreliable, mixed, reliable	Rótulos de informações básicas no nível da fonte de 7 locais de avaliação diferentes.	1,12M

oriundos destas plataformas, apresentamos uma visão geral da arquitetura por trás de um coletor do WhatsApp e de Telegram. As primeiras etapas para a coleta de dados em ambos aplicativos são a criação de um perfil (ou persona), a busca e a entrada nos grupos identificados como relevantes para a coleta. Na Figura 1.3 apresentamos um esquema de como esse processo é realizado desde a autenticação do celular, a busca em plataformas sociais até a entrada nos grupos. É importante mencionar que, para o primeiro passo, da criação da conta, é necessário a ativação do *SIM Card* com um número de celular válido e um *smartphone* conectado à Internet. Com o celular preparado, podemos instalar tanto o aplicativo do WhatsApp como do Telegram (entre outros mensageiros) e fazer a autenticação dele com o número dedicado para a coleta. Após a verificação, cria-se um perfil no aplicativo baseado no tema da coleta.

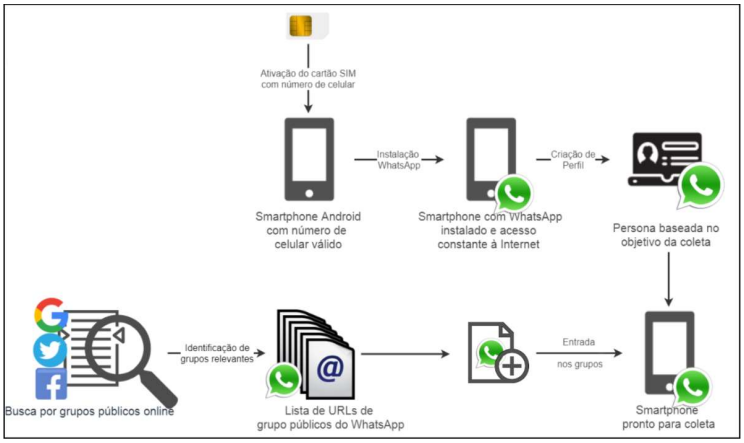


Figura 1.3. Visão geral sobre o processo de autenticação e entrada dos grupos.

Especificamente para o WhatsApp, outra etapa necessária é a preparação da conta de WhatsApp Web. Essa será a versão do WhatsApp utilizada por um dos coletores para acessar os dados através de *scripts* executados no servidor. Assim, é necessário fazer login no WhatsApp Web também por um computador na primeira vez que a ferramenta desenvolvida neste projeto for utilizada. A Figura 1.3 mostra um fluxograma de como se conectar na versão do WhatsApp Web. A partir do fluxograma, podemos observar que existe uma etapa manual que exige o uso do celular com o WhatsApp que se deseja coletar para fazer a leitura do QR Code exibido na página para conexão. Essa etapa é necessária apenas na primeira vez de configuração da coleta. Uma vez conectado, às demais instâncias de coleta já estarão com a conta devidamente logada. Entretanto, é válido destacar que eventualmente, a conta WhatsApp pode "deslogar-se" sozinha do navegador devido a atualizações do próprio WhatsApp, por exemplo, ou até mesmo caso o perfil do *browser* (e.g., Firefox) seja apagado da memória. Logo, esse processo poderá ser repetido sempre que necessário (i.e., quando a conta for desconectada do navegador). A seguir, apresentamos detalhes de cada uma das etapas.

Etapa 1 – Criação de uma Persona. Coletar dados do WhatsApp ou do Telegram requer uma interação muito maior com os outros usuários comparado a outras plataformas como Instagram, YouTube ou Twitter. É preciso efetivamente participar dos grupos no mesmo nível dos outros membros, não podendo somente observá-los a distância. Com isso, é necessária uma cautela maior para respeitar os Termos de Serviço do WhatsApp e também proceder com uma observação mais ética do projeto. Por isso, durante a criação do perfil para acompanhamento dos grupos públicos, exploramos o conceito de persona. Uma persona é uma representação fictícia e objetiva de um perfil de usuário, cuja idealização é baseada em pesquisas com usuários, observação de interesses, desejos e necessidades. Este é um mecanismo frequentemente utilizado durante o processo de desenvolvimento de um *software*. Aqui, esta abordagem foi utilizada para a coleta dos grupos de plataformas de mensagens instantâneas, uma vez que ela é capaz de sintetizar a aparência de um usuário convencional dos aplicativos alvo.

Para criação de uma persona para os grupos de WhatsApp ou Telegram, primeiro é necessário responder algumas perguntas relacionadas à sua personalidade, vida e o uso do aplicativo, tendo em mente os grupos de interesse para observação. As perguntas utilizadas são: *(a)* Qual o gênero da pessoa?; *(b)* Qual sua idade?; *(c)* A pessoa é casada,

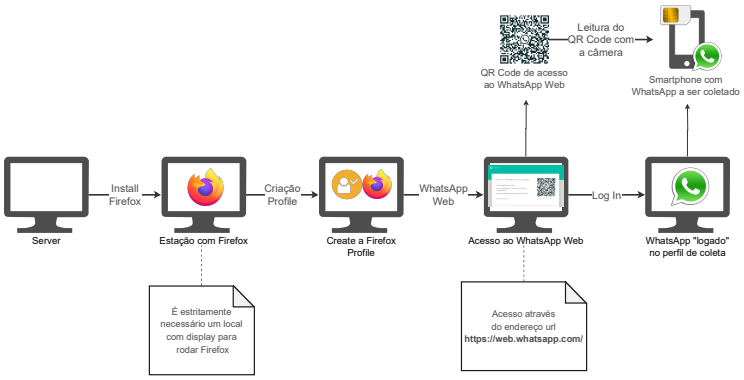


Figura 1.4. Fluxograma de conexão com o WhatsApp Web.

ou está em um relacionamento?; *(d)* A pessoa tem filhos?; *(e)* A pessoa mora sozinha, ou divide a casa? Com quem?; *(f)* Qual seu grau de escolaridade (e.g., ensino fundamental completo, ensino médio completo)?; *(g)* A pessoa trabalha? Qual o seu trabalho?; *(h)* Por que a pessoa utiliza a plataforma?; *(i)* Por quanto tempo ela utiliza o aplicativo de mensagens?; *(j)* A pessoa acessa pelo celular, computador, ou tablet?; e, *(k)* Por quais grupos ela se interessa?

A partir dessas questões é possível imaginar um usuário alvo da coleta a ser realizada, com uma personalidade que se adéque aos grupos desejados, e não seja identificada como um perfil não autentico pelos outros membros do grupo. Ademais, para a criação de personas é sugerido o uso de imagens sem restrições de uso comercial, ou imagens de pessoas que não existam (i.e., criadas com o auxílio de ferramentas de inteligência artificial). Seguindo o ideal da persona desenvolvida, além de uma metodologia já consolidada de etnografia virtual, é criado também um ponto de referência fixado na realidade dos usuários, mantendo a seleção e entrada nos grupos públicos em acordo com a proposta do projeto. Além disso, seguir a persona é uma forma de garantir que o perfil respeite os Termos de Serviço do WhatsApp, evitando, entre outras coisas, o banimento da plataforma.

Etapa 2 – Busca de Grupos Públicos Relevantes. A segunda etapa, que pode acontecer paralelamente à anterior, está relacionada a busca por grupos de interesse que serão coletados. Tanto o WhatsApp, o Telegram como algumas outras plataformas de mensagens instantâneas possuem grupos públicos de conversa nos mais variados temas. Esses grupos públicos são criados por usuários e compartilhados na Web através de URLs de convites. A ideia é que outras pessoas possam participar livremente das conversas. Como o objetivo desses usuários é divulgar seus grupos para agregar mais membros, é possível achar facilmente na Internet, como em outras plataformas sociais ou ainda em *websites* de buscas, dezenas e até centenas de grupos de tópicos variados. Existem até *websites* que funcionam como repositório próprios de grupos de WhatsApp separados por categorias[31].

Uma vez definido um foco para coleta e a persona que participará dos grupos, precisamos então levantar uma lista dessas URLs de grupos púbicos considerados relevantes para a coleta. Por exemplo, grupos públicos relacionados à política ou então com informações relacionadas à pandemia de COVID-19. De posse da lista de grupos, ainda é necessário "entrar" nesses grupos. Esse processo pode ser feito de forma manual, aces-

[31] `https://gruposdezap.com/`

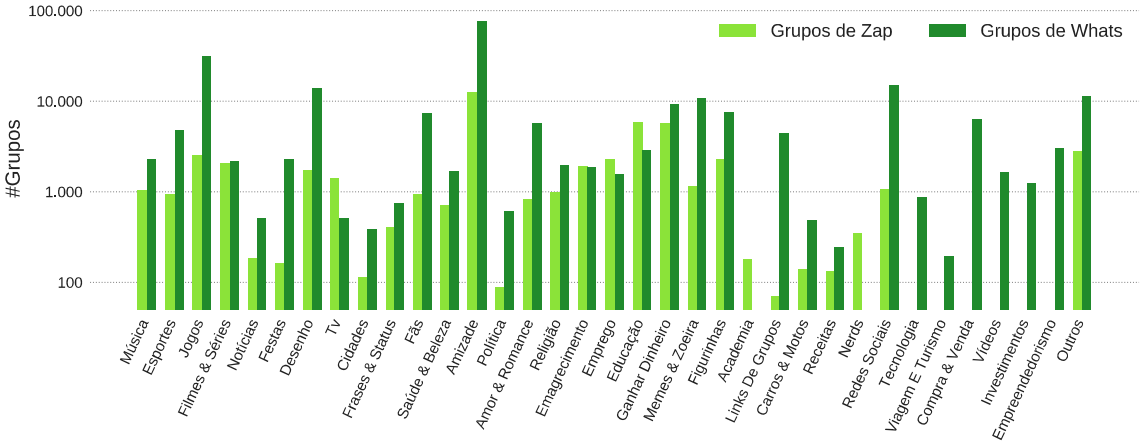


Figura 1.5. Quantidade de grupos criados por categoria nos repositórios online.

sando o aplicativo através do celular da coleta e ir clicando grupo a grupo e dando o "*Join*" naqueles grupos, ou, quando a lista é muito extensa, também é possível automatizar esse processo. No Telegram, através da API do sistema, podemos ir "entrando" em grupos de interesse, dado seu ID. Já no WhatsApp, que não dispõe de uma API, isso pode ser realizado por meio de um *script* Python com uso do *Selenium*[32], por exemplo, que faça uso do WhatsApp Web para simular o processo manual que um usuário faria indo por cada link de grupo público e "participando" dele.

Note porém, que a existência da URL de um grupo público não garante o acesso a ele. As URLs podem ser revogadas pelos administradores dos grupos, restringindo o acesso e, mesmo após a entrada em um grupo, é possível que a persona criada seja banida (expulsa) dele de acordo com a vontade dos administradores ou baseado em regras específicas definidas pelo grupo.

Em um trabalho desenvolvido recentemente pelo nosso grupo de pesquisa [Kansaon et al., 2022], nós mapeamos de forma geral as categorias de grupos de WhatsApp criados no Brasil. Para isso avaliamos dois grandes *websites* repositórios de grupos e também aqueles compartilhados no Twitter e Facebook. Com isso descobrimos as principais categorias e tópicos de grupos públicos criados no Brasil. Esta investigação nos ajuda a planejar melhor novas pesquisas que envolvem este tipo de conteúdo, uma vez que possibilita entender melhor o ecossistema do WhatsApp e que tipo de conteúdo é criado nele.

Na Figura 1.5 apresentamos as principais categorias de grupos criados no Brasil de dois dos principais repositórios de grupos no país, a saber: "Grupos de Zap" e "Grupos de Whats". É interessante observar o uso dos grupos pelos usuários, com a categoria Amizade sendo a mais popular. Além disso, outras categorias também se destacam, como a categoria de Jogos (33.954); Desenho (15.603), considerando desenhos animados e animes; Memes e Zoeira (11.846), com compartilhamento de piadas, memes e sátiras; Esportes (5.688), composto na maioria por grupos de futebol, e de Redes Sociais (16.192), com vários grupos com foco em outras plataformas como YouTube, Facebook, TikTok, Twitter e Instagram. Uma grande parte dos grupos desta categoria são sobre ganhar seguidores e curtidas em outras plataformas digitais.

Ao compartilhar grupos nesses ambientes, os usuários fornecem informações de

[32]`https://selenium-python.readthedocs.io/`

(a) Twitter (b) Facebook

Figura 1.6. Nuvem de termos das mensagens contendo grupos de WhatsApp.

contexto que descrevem o tema dos grupos. Assim, também analisamos os tópicos dos grupos compartilhados no Twitter e no Facebook. As nuvens de termos da Figura 1.6 de ambas as plataformas, revela diferenças marcantes entre elas. No Twitter, termos com conotação sexual são mais frequentes, enquanto no Facebook prevalecem palavras religiosas. As mensagens também fazem uso de termos metalinguísticos para convidar pessoas a ingressar em grupos do WhatsApp. Muitas palavras referem-se a outras plataformas sociais, como Telegram, TikTok e Instagram, indicando uma potencial relação entre elas. Além disso, foram encontrados termos relacionados à venda de produtos e serviços, sugerindo que o WhatsApp é muitas vezes usado como canal para negociações diversas.

Etapa 3 – Extração, Armazenamento e Processamento dos Dados. Finalmente, a Figura 1.7 apresenta uma visão geral de todos os passos necessários para a coleta de dados dessas plataformas, bem como os objetos envolvidos durante cada etapa do processo de execução: *(1)* O celular e o computador configurados para realizar a coleta, ambos conectados ao WhatsApp/Telegram; *(2)* Um *script* observa os dados recebidos nos grupos monitorados e extrai as mensagens; *(3)* As mensagens são estruturadas e salvas em JSON; *(4)* Os arquivos de mídia (i.e., imagem, vídeo e áudio) são baixados e salvos no servidor; *(5)* Outro *script* é responsável por analisar esses arquivos e agrupar as mensagens por similaridade; *(6)* As mensagens agrupadas são salvas em JSON contendo todas vezes que elas foram compartilhadas.

É importante mencionar que propusemos e adotamos uma estratégia mais pragmática para obtenção de metadados relacionados a cada grupo de interesse, coletando as informações imediatamente após o ingresso no grupo descoberto. Ademais, abordamos a metodologia exata para cada plataforma de mensagens da seguinte maneira. Para subconjunto dos grupos monitorados, complementamos os metadados básicos do grupo com detalhes sobre a estrutura e atividade dentro deles. Como a metodologia para esta etapa difere drasticamente para cada plataforma, abaixo, apresentamos detalhes relativos à obtenção de dados dos grupos para cada uma delas (i.e., WhatsApp e Telegram).

WhatsApp. Com a falta de suporte de API para o WhatsApp, há uma dependência da interface Web do WhatsApp para ingresso nos grupos e coleta das informações de interesse. Neste contexto propusemos e desenvolvemos uma abordagem em Python utilizando *Selenium* para salvar os dados enviados nos grupos da conta monitorada. A partir desse código implementado, temos acesso a: *(i)* mensagens dos grupos (o WhatsApp dá acesso apenas às mensagens enviadas após a data de ingresso); *(ii)* números de telefone de todos os membros do grupo; e *(iii)* data de criação do grupo.

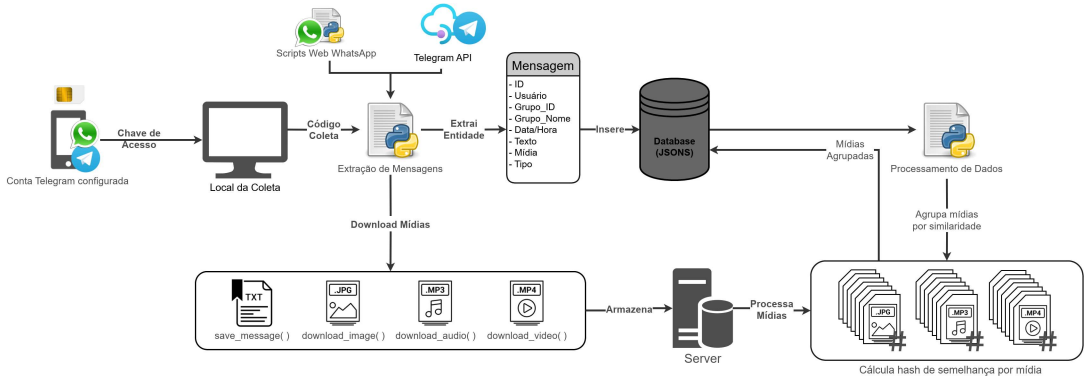


Figura 1.7. Visão geral da arquitetura da coleta de dados em aplicativos de mensagem instantânea.

Telegram. Por outro lado, obter dados dos grupos do Telegram é uma tarefa menos complexa em comparação ao WhatsApp devido sua API [33]. Logo, após ingressar nos canais e grupos, temos mais informações e também acesso a todas as mensagens, incluindo aquelas anteriores ao nosso ingresso. Para cada grupo, coletamos: *(i)* mensagens dos grupos (todas as mensagens desde a criação do grupo/canal); *(ii)* data de criação do grupo; e, *(iii)* perfis de usuário dos membros do grupo.

Aqui, vale notar uma etapa crucial para o funcionamento do sistema: o agrupamento de conteúdo semelhante. Em outras plataformas sociais, por exemplo, quando nos referimos a coleta de uma mensagem postada no Facebook ou Twitter, as informações disponibilizadas comumente contém metadados relacionados (e.g., número de curtidas ou compartilhamentos da referida mensagem dentro da plataforma). Porém, no WhatsApp, cada mensagem, cada imagem, vídeo ou áudio é postada de forma independente e isolada. Para que seja possível, por exemplo, mensurar quantas vezes um determinado conteúdo foi compartilhado na plataforma, é necessária a execução de um processo específico que envolve o processamento, rastreamento, agrupamento e contagem de um determinado conteúdo. Portanto, para saber que uma única mensagem foi compartilhada mais de uma vez, precisamos fazer esse processamento de forma manual. Este processo é apresentado em detalhes no fluxograma da Figura 1.8. Aqui vemos como as mensagens isoladas, mas que correspondem ao mesmo conteúdo, são agrupadas pela nossa metodologia. Quando dois usuários compartilham a mesma imagem, por exemplo, elas são salvas de forma isolada pelo coletor. Para traçar essa disseminação de um conteúdo de mídia através do WhatsApp realizamos um agrupamento dos conteúdos por similaridade usando um sistema de *hashes*. Para os arquivos multimídia recebidos, então, o *script* de processamento realiza os seguintes passos: *(1)* baixa os arquivos anexados a uma mensagem durante a coleta; *(2)* calcula a *hash* de cada um desses arquivos[34]; *(3)* cria-se um dicionário *hash* das mídias coletadas, de forma que duas mídias idênticas possuem a mesma hash, podendo assim identificar mensagens distintas que compartilharam o mesmo conteúdo; *(4)* agrupa, finalmente, as mensagens com todas as ocorrências em que ela foi compartilhada, armazenando dados como total de vezes que ela foi enviada, quais foram os usuários e grupos que compartilharam este conteúdo e datas de postagem.

Por fim, as mensagens coletadas especificamente no WhatsApp foram utilizadas

[33] `https://core.telegram.org/method/channels.joinChannel`

[34] Usando o *checksum* MD5 do arquivo para áudios e vídeos e uma *perceptual hash* para imagens.

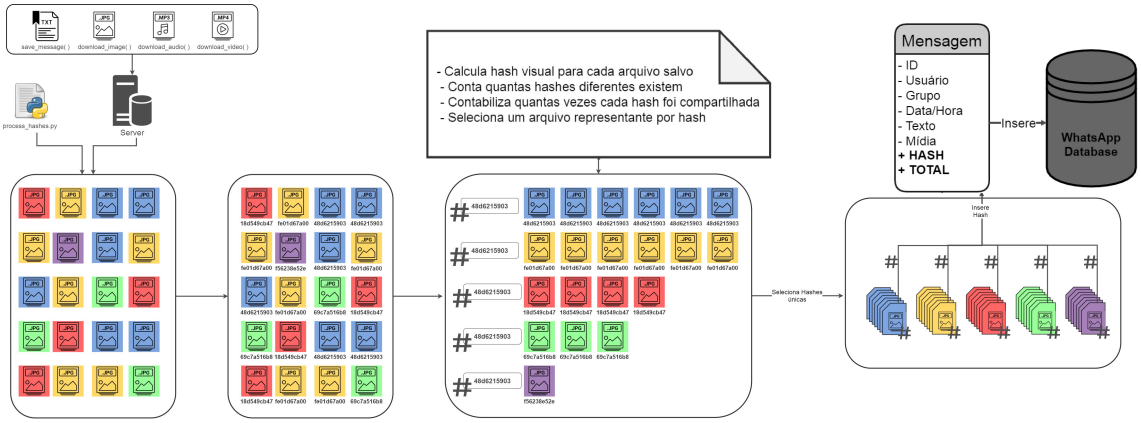


Figura 1.8. Abordagem de agrupamento do conteúdo por semelhança de *hash*.

para a construção de um conjunto de dados disponível em `https://zenodo.org/record/3779157` que tem sido amplamente utilizado na literatura para entendimento do fenômeno da desinformação bem como para proposição de abordagens que possam ser úteis para contenção do problema nesses ambientes [Reis et al., 2020a; Reis and Benevenuto, 2021; De Angeli and Reis, 2022].

1.5. Entendimento/Caracterização da Desinformação

Embora o fenômeno da desinformação em si não seja um problema novo[35], recentemente, alguns esforços estão emergindo com o objetivo de melhor compreendê-lo em diferentes plataformas digitais, como Twitter, WhatsApp, Facebook, Telegram, etc. Particularmente, Vosoughi et al. [Vosoughi et al., 2018] mostra que notícias contendo desinformação tendem a se espalhar mais rapidamente do que notícias contendo histórias verdadeiras. Lazer et al. [Lazer et al., 2018] convocam uma força-tarefa interdisciplinar para abordar esse problema complexo. No entanto, existem algumas características inerentes às próprias plataformas digitais que contribuem para a disseminação deste tipo de conteúdo nesses ambientes. Em suma, estes esforços para entendimento do fenômeno estão focados em diferentes aspectos, tais como as características do conteúdo, dinâmica de propagação (e.g., compartilhamentos), dentre outros.

Assim, nesta seção, serão discutidas as principais estratégias empregadas para entendimento do fenômeno da desinformação em plataformas digitais e meios de investigar sua disseminação através de diferentes plataformas, como, por exemplo, no WhatsApp, onde o nosso grupo de pesquisa se destaca como um dos pioneiros no Brasil [Resende et al., 2019]. Neste contexto, a natureza fechada do aplicativo, juntamente com a facilidade de compartilhamento de informações em grupos de grande escala, tornam o WhatsApp único entre outras plataformas, onde mensagens criptografadas anônimas podem se tornar virais, alcançando vários usuários em um curto intervalo de tempo. A sensação de pessoalidade e a imediatidade das mensagens foram amplamente abusadas para espalhar rumores infundados e criar campanhas de desinformação durante as eleições recentes no Brasil e na Índia [Reis et al., 2020a]. Apesar do esforço para combater o problema, não há evidências até o momento sobre a real eficácia de tais restrições. Aqui, investigamos e

[35] `http://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535`

propomos uma série de estratégias para, por exemplo, investigar e entender a eficácia dessas medidas na disseminação de desinformação circulando no WhatsApp. Através do uso de um modelo epidemiológico com dados reais coletados do WhatsApp no Brasil, Índia e Indonésia, avaliamos o impacto da viralidade nesse tipo de rede [Melo et al., 2019b]. Conforme mencionado anteriormente, discutiremos estas, e outras abordagens exploradas na literatura para entendimento/caracterização do problema em plataformas digitais, nesta seção.

1.5.1. Propagação de (Des)Informação no WhatsApp

Existem algumas características-chave que tornam o WhatsApp único em relação a outras plataformas sociais. Em primeiro lugar, o WhatsApp permite a conexão entre pessoas com interesses semelhantes por meio de grupos de bate-papo. Esses grupos têm um limite de 256 usuários e podem ser privados ou públicos. No caso dos grupos privados, novos membros precisam ser adicionados por um membro que assume o papel de administrador do grupo. Para os grupos públicos, o acesso é feito por meio de links de convite que podem ser compartilhados com qualquer pessoa ou disponibilizados na Web. Esses grupos públicos são frequentemente usados para discutir hobbies, paixões e também tópicos específicos, como saúde, educação e política.

Embora a maioria dos grupos seja privada e formada por pessoas com algum relacionamento social (por exemplo, família, amigos, colegas de trabalho), os grupos públicos têm sido uma característica catalisadora para a difusão de informações neste ecossistema: a maioria de seus membros são desconhecidos entre si. Isso é evidente em países como o Brasil, onde uma pesquisa relatou que 76% dos usuários do WhatsApp fazem parte de grupos, 58% participam de grupos com pessoas desconhecidas e 18% desses grupos discutem política [Newman et al., 2019]. Desta forma, grupos públicos de WhatsApp podem atuar como um atalho para que a informação percorra partes distantes da estrutura da rede social subjacente por meio de laços fracos, ampliando e acelerando a disseminação de informações [Bakshy et al., 2012]. Além disso, o aplicativo facilita ainda mais a propagação de informação uma vez que possui duas funções de compartilhamento: o *broadcast*, em que uma lista de contatos pode ser criada para enviar mensagens para até 256 contatos (usuários ou grupos) de uma só vez, e o *forward*, em que uma única mensagem recebida pode ser prontamente encaminhada para outros contatos (usuários ou grupos). Essas características permitem que a mensagem percorra rapidamente longas distâncias pela rede. Por outro lado, a criptografia ponta-a-ponta implementada pelo sistema dificulta a identificação da origem e o rastreamento da propagação destas mensagens.

Devido a essas peculiaridades, o WhatsApp tem gerado uma controvérsia relacionada às suas características de anonimato e viralidade. Esse conflito se deve ao fato de que podemos enxergar o WhatsApp de duas maneiras diferentes: como uma empresa de tecnologia que oferece um serviço de mensagens seguras e privadas e, ao mesmo tempo, como uma plataforma de mídia de comunicação em massa. Por um lado, ele garante o anonimato, privacidade e a segurança do usuário por meio da criptografia de dados. Como meio de comunicação em massa, ele transmite informações e dissemina conteúdo em grande escala em um curto espaço de tempo. Assim, mensagens enviadas anonimamente alcançam rapidamente milhares de pessoas sem qualquer regulação ética ou legal desse conteúdo disseminado, promovendo, por exemplo, campanhas de desinformação. A

disseminação massiva de (des)informação e boatos [Arun, 2019] levou, inclusive, a pedidos tanto dos governos nacionais da Índia e posteriormente do Brasil [Melo et al., 2019b] para o WhatsApp para que eles alterassem recursos tentando minimizar os danos causado pela plataforma sendo usada para espalhar desinformação em grande escala. Como consequência, o WhatsApp implementou restrições na forma como as mensagens são encaminhadas[36], reduzindo o limite de encaminhamento de conteúdo para um máximo de usuários/grupos simultâneos. No entanto, não existem estudos que investiguem o impacto dessas limitações ou se os números escolhidos pela empresa são suficientes para barrar eficientemente a propagação de conteúdo viral.

Nesta seção, avaliamos a dinâmica da propagação da (des)informação em uma rede de grupos públicos do WhatsApp. Mais especificamente, investigamos a anatomia dessa plataforma para compreendermos suas peculiaridades, bem como respondermos à pergunta de como as ferramentas de encaminhamento contribuem para a viralidade de (des)informação. Além disso, se as limitações impostas no sistema são realmente capazes de evitar a disseminação de conteúdo. Por fim, também propomos uma estratégia para medição dessa propagação em rede do WhatsApp e discutimos algumas soluções de como o problema da disseminação em larga escala pode ser contido/combatido.

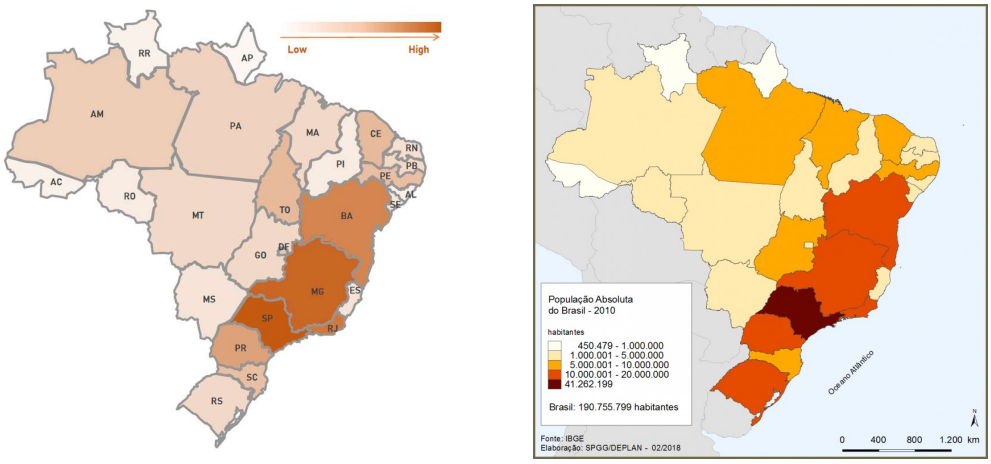
1.5.1.1. A Estrutura de Rede do WhatsApp

Iniciamos esta análise exploratória com a coleta de dados do WhatsApp para obter informações sobre a estrutura da rede e suas características. Os dados utilizados foram coletados por [Resende et al., 2019] durante o período eleitoral brasileiro entre 16 de setembro e 5 de outubro de 2018. Os dados analisados se referem especificamente às imagens compartilhadas em 364 grupos políticos públicos; adquiridos a partir de uma busca na Web utilizando de uma lista de palavras-chave relacionadas à política e notícias, juntamente com o link `chat.whatsapp.com`, para identificar e coletar grupos públicos do WhatsApp por meio de pesquisas no Google, Twitter e Facebook. Esta coleta, em particular, foi realizada utilizando a ferramenta WebWhatsAppAPI, que permite a extração de mensagens por meio da versão Web do WhatsApp. Além disso, complementamos nosso conjunto de dados com conteúdos semelhantes de grupos públicos de WhatsApp da Índia e da Indonésia em uma coleta em parceria com autores de [Garimella and Tyson, 2018].

Após a coleta dos dados, procedemos com uma caracterização inicial dos dados e a reconstrução da estrutura de rede dos grupos públicos de WhatsApp que estes grupos constituíam a fim de entender as propriedades básicas desta plataforma. O objetivo é identificar comportamentos e vieses dos dados que diferenciam a rede de grupos do WhatsApp de outras redes sociais, como Facebook e Twitter.

Além disso, exploramos a construção de um modelo generativo para criar um grafo que representasse adequadamente a rede de usuários do WhatsApp, levando em consideração suas características peculiares, como a presença de grupos de conversa. Propusemos uma divisão da rede em dois tipos de grupos: orgânicos e artificiais. Essa distinção permitiu uma análise mais aprofundada da estrutura da rede e das interações entre os grupos e seus membros.

[36] `blog.whatsapp.com/10000647/More-changes-to-forwarding`



(a) Usuários do WhatsApp por estado

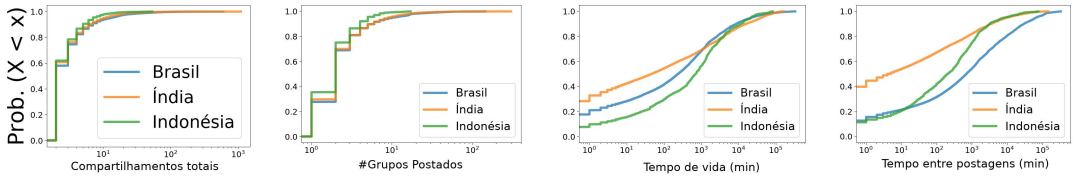
(b) População absoluta do Brasil.
Fonte: IBGE/Censo Demográfico 2010

Figura 1.9. Comparação populacional entre a geolocalização dos usuários do *WhatsApp* baseado no código DDD de telefone e população do Brasil.

Nossa primeira análise consiste em identificar os padrões geográficos contidos nos dados. Como cada usuário está associado a um número de telefone, é possível localizar geograficamente os usuários através de seu código DDD. A Figura 1.9 apresenta o mapa coroplético das localizações dos 10 mil usuários coletados distribuídos pelas unidades federativas do Brasil. Cores mais escuras indicam maior concentração de usuários naquele estado. Os estados com maior número de usuários incluem SP (1322), MG (1116), RJ (992) e BA (905). Esses valores confirmam o viés geográfico da distribuição real 1.9(b) da população sobre o território brasileiro, o que demonstra que os dados possuem uma amostra de usuários semelhante à distribuição populacional brasileira. Entretanto, podemos observar maiores discrepâncias para os estados do RS (275), que aparece subrepresentado, TO (479) e DF (380), que possuem proporcionalmente mais usuários do que a distribuição de população. Como se tratam de dados políticos, relacionamos esse aumento no DF à sua proximidade da vida política do país. Já Tocantins aparece com uma concentração também acima do normal devido a uma grande quantidade de grupos monitorados referentes àquele estado.

Uma vez que coletamos e identificamos as ocorrências dos conteúdos compartilhados nesses grupos, podemos observar a cobertura e a dinâmica de propagação dessas imagens em nossos dados. Para avaliar métricas de propagação ao longo do tempo e cobertura, consideramos apenas as imagens que foram compartilhadas pelo menos duas vezes, pois não podemos observar o efeito da propagação de imagens que são postadas apenas uma vez. Esse conjunto consiste em 2.384 imagens na Indonésia, 103.031 imagens no Brasil e 44.731 imagens na Índia, o que representa aproximadamente 20% das imagens em cada país. Embora quase 80% das imagens no WhatsApp tenham sido postadas apenas uma vez, existem algumas imagens muito populares que foram amplamente compartilhadas e atingiram vários grupos.

Primeiro, calculamos o número total de compartilhamentos de cada imagem e em quantos grupos elas apareceram. As Figuras 1.10(a) e 1.10(b) mostram a função de distribuição cumulativa (CDF) do número total de compartilhamentos e do número de grupos distintos em que cada imagem apareceu. É possível observar que existem algumas



(a) Compartilhamentos (b) Grupos (c) Tempo de vida (d) Intervalo entre postagens

Figura 1.10. CDF de cobertura de compartilhamento e métricas de dinâmica de tempo de imagens compartilhadas pelo menos duas vezes no WhatsApp.

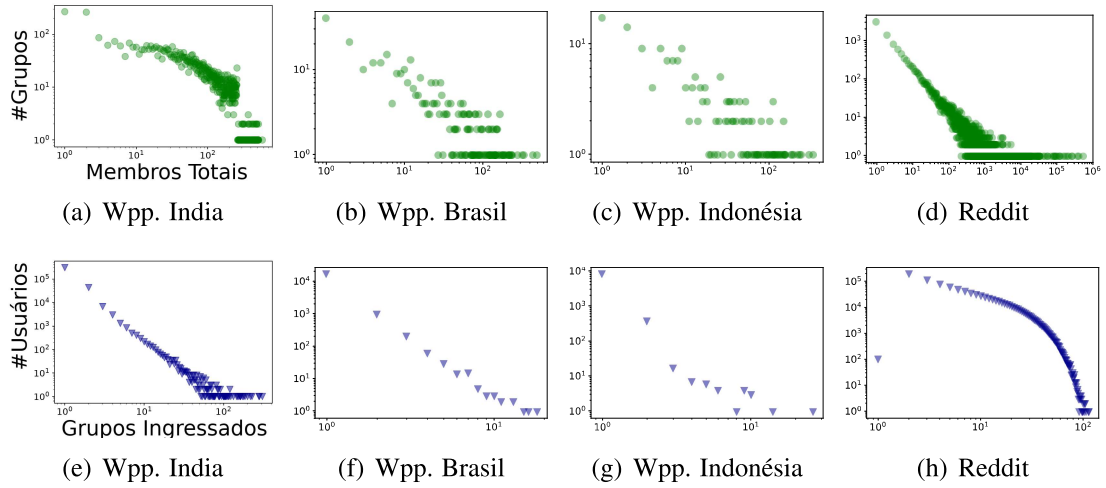
imagens muito populares que foram amplamente compartilhadas mais de 500 vezes no Brasil e mais de mil vezes na Índia, e alcançaram mais de 100 grupos em ambos os países. Embora a maior parte das imagens tenha sido compartilhada poucas vezes, isso mostra que o WhatsApp pode ser usado não apenas para conversas particulares, mas também como um meio de comunicação em massa com potencial de viralização de seu conteúdo.

Além de analisar a propagação das imagens no WhatsApp, também analisamos sua "vida útil"na Figura 1.10(c). A vida útil é determinada pela diferença entre a última e a primeira ocorrência da imagem em nosso conjunto de dados. Em resumo, embora a maioria das imagens (80%) dure no máximo 2 dias, existem imagens no Brasil e na Índia que continuaram a aparecer mesmo após 2 meses da primeira aparição (100.000 minutos). Também podemos observar que a maioria (60%) das imagens é postada antes de 1.000 minutos após sua primeira aparição. Além disso, no Brasil e na Índia, cerca de 40% dos compartilhamentos foram feitos após um dia de sua primeira aparição e 20% após uma semana. Em uma análise mais detalhada, na Figura 1.10(d), mostramos a distribuição dos "tempos entre eventos"entre postagens da mesma imagem. Observamos que o tempo entre eventos das imagens na Índia é muito mais rápido do que no Brasil e na Indonésia, ou seja, mais de 50% das postagens são feitas em intervalos de 10 minutos ou menos, enquanto apenas 20% dos compartilhamentos foram feitos nesse mesmo intervalo de tempo no Brasil e na Indonésia. Analisamos manualmente as razões por trás do curto período de tempo entre as postagens e descobrimos que nos dados da Índia há um comportamento automatizado semelhante a spam em comparação com o Brasil e a Indonésia.

Esses resultados sugerem que o WhatsApp é uma rede muito dinâmica e a maior parte do seu conteúdo de imagens é efêmera, ou seja, as imagens geralmente aparecem e desaparecem rapidamente. A arquitetura linear do sistema de chat de conversas dificulta que um conteúdo antigo seja revisitado, mas há alguns que permanecem na rede por mais tempo, disseminando-se ao longo de semanas ou mesmo meses.

Na Figura 1.11, mostramos a distribuição de grupos por usuário e usuários por grupo. Para comparar as peculiaridades do WhatsApp com outras plataformas populares, também usamos dados da rede Reddit, modelando os subreddits como grupos e os usuários como membros. Observe que, embora o Reddit tenha a mesma característica de grupos, queremos avaliar recursos específicos do WhatsApp que levam a estruturas de rede muito diferentes. O limite de 256 membros nos grupos é um elemento determinante na rede, capaz de limitar o tamanho do grupo, principalmente na Índia (Figura 1.11(a)),

onde existem mais de 300 mil usuários e mais de 5 mil grupos.[37] Por outro lado, no Reddit, onde não há limite, é possível ver que o tamanho do grupo pode ser tão grande quanto 10^5 membros, o que cria grandes concentrações de usuários. Como ambas as plataformas não têm limite para o número de grupos aos quais os usuários podem aderir, esperávamos não ver diferenças no número total de grupos dos quais os usuários participam. No entanto, observe que no Reddit, a distribuição tem um decaimento exponencial, com um limite de aproximadamente 100 grupos. Por outro lado, todas as curvas do WhatsApp são semelhantes, seguindo uma curva de lei de potência bem comportada, o que naturalmente gera uma variância maior. Observe que na Índia temos usuários que participaram de mais de 300 grupos.



(a) Wpp. India (b) Wpp. Brasil (c) Wpp. Indonésia (d) Reddit

(e) Wpp. India (f) Wpp. Brasil (g) Wpp. Indonésia (h) Reddit

Figura 1.11. Distribuições do número de membros por grupo e total de grupos por usuário no WhatsApp (Wpp) e no Reddit.

Em seguida, investigamos a estrutura de rede dos grupos públicos do WhatsApp e comparamos suas características com outras redes sociais reais e sintéticas. Para reconstruir a rede a partir dos grupos do WhatsApp coletados, criamos um grafo em que conectamos dois grupos se eles compartilham um usuário em comum. Embora o WhatsApp seja um aplicativo de chat pessoal criptografado, a possibilidade de criar grupos públicos permite que vários usuários distantes socialmente se conectem uns aos outros por meio da rede, formando uma estrutura social complexa capaz de fluir grandes volumes de informações.

Na Figura 1.12, mostramos essas redes para os três países, onde cada nó representa um grupo e as arestas conectam os nós que possuem membros em comum. O tamanho do nó é proporcional ao número de membros do grupo. Colorimos os nós de acordo com sua comunidade nesse grafo, seguindo o algoritmo de modularidade proposto por [Blondel et al., 2008]. Observe que em todos os gráficos há um componente conectado principal evidente e outros agrupamentos de grupos. Além disso, note que alguns grupos se posicionam como pontes e *hubs*, conectando diferentes comunidades na estrutura da rede.

[37]Em nossos dados, alguns grupos têm mais de 256 membros, porque nossos dados são uma captura temporal e os membros podem sair e entrar nos grupos durante esse tempo.

(a) Brasil (b) Índia (c) Indonésia

Figura 1.12. Rede de grupos públicos do WhatsApp para cada país. Cada nó é um grupo e as arestas representam membros em comum.

Esses resultados demonstram que o WhatsApp apresenta uma estrutura de rede social semelhante a outras redes sociais estudadas. A existência de um componente conectado principal indica que há uma grande interconectividade entre os grupos, permitindo a disseminação de informações de forma eficiente. Além disso, a presença de agrupamentos de grupos indica a existência de comunidades ou temas específicos dentro da rede do WhatsApp. Ao identificar grupos que atuam como "pontes" e "hubs", podemos inferir que esses grupos desempenham um papel fundamental na conectividade entre diferentes comunidades na rede. Eles podem ser responsáveis por ampliar o alcance das informações, permitindo que elas se espalhem entre grupos distintos. Essa característica do WhatsApp como uma plataforma de comunicação em grupo pode ser explorada para fins de disseminação de informações, seja de maneira positiva ou negativa.

Em suma, o estudo da estrutura de rede dos grupos do WhatsApp nos permite entender melhor como as informações são compartilhadas e propagadas dentro da plataforma. Isso pode ter implicações significativas para a compreensão do papel do WhatsApp como meio de comunicação e para o desenvolvimento de estratégias de gerenciamento de informações na era digital.

A seguir, comparamos as características destas redes de grupos do WhatsApp com outras redes: *(i)* gráficos gerados aleatoriamente usando o modelo Barabasi-Albert de rede livre de escala, o modelo Erdős–Rényi, o modelo de mundo pequeno [Watts and Strogatz, 1998] e o modelo de rede *Forest Fire* [Leskovec et al., 2005], para os quais usamos o mesmo número de nós no conjunto de dados da Índia para criar uma rede comparável; *(ii)* a rede de subreddits do Reddit [Olson and Neal, 2015]; e *(iii)* a rede do Flickr [McAuley and Leskovec, 2012], que, diferentemente das redes de grupos do WhatsApp e do Reddit, representa a rede de imagens compartilhadas pelos usuários na plataforma.

Os resultados são apresentados na Tabela 1.2. Observamos que o WhatsApp compartilha características comuns com outras redes sociais do mundo real: alto coeficiente de agrupamento, maior componente conectado gigante e pequeno comprimento médio do caminho, que são propriedades típicas de uma rede social. O WhatsApp também apresenta um coeficiente de Pearson mais alto, o que significa que os nós tendem a se conectar com outros nós que possuem valores de grau semelhantes. Em análises de epidemia, isso pode ajudar a entender a propagação de infecções na rede, pois uma campanha de desin-

Tabela 1.2. Métricas de rede para o WhatsApp em comparação com outras redes.

	#Nós	#Arestas	Grau Médio	Coeficiente Clusterização	Diâmetro	APL*	Densidade	LCC**	Coeficiente Pearson
Wapp. India	5,839	407,081	69.71	0.59	11	3.17	0.0239	92.6%	0.295
Wapp. Brazil	414	1,400	6.76	0.32	8	3.19	0.0164	65.2%	0.346
Wapp. Indonesia	217	699	6.44	0.38	9	3.09	0.0298	55.3%	0.290
Bar.-Albert	5,839	792,300	271.38	0.10	3	1.95	0.0465	100%	0.008
Erdos-Renyi	5,839	1,534,952	525.76	0.09	2	1.91	0.0901	100%	-0.001
Smallworld	5,839	604,250	206.97	0.34	3	1.98	0.0355	100%	0.007
ForestFire	5,839	12,930	4.43	0.42	17	5.25	0.0008	100%	-0.066
Reddit	15,122	4,520,054	597.81	0.82	6	2.03	0.0395	99,8%	-0.045
Flickr	105,938	2,316,948	43.74	0.09	9	4.8	0.0004	99.8%	0.247

*Comprimento Médio do Caminho (do Inglês, *Average Path Length)*
**Maior Componente Conectado (do Inglês, *Largest Connected Component)*

formação direcionada a grupos de alto grau provavelmente se espalhará para outros nós de alto grau.

Essas semelhanças indicam que o WhatsApp possui uma estrutura de rede social robusta, compartilhando características com outras redes sociais populares. Isso destaca a importância do WhatsApp como uma plataforma de comunicação e troca de informações entre os usuários. Compreender a estrutura dessa rede pode ser útil para analisar a disseminação de informações e desenvolver estratégias para lidar com a propagação de conteúdos falsos ou prejudiciais.

Ao compreender a estrutura da rede do WhatsApp e suas particularidades, podemos avançar nas análises relacionadas à propagação de (des)informação e entender como as ferramentas de encaminhamento contribuem para a viralidade de conteúdo. Essas informações são cruciais para o desenvolvimento de estratégias eficazes de combate à disseminação de desinformação em larga escala no WhatsApp e em outras plataformas de comunicação.

1.5.1.2. Propagação de Informação através do Modelo Suscetível-Exposto-Infectado

Com a rede pronta, precisamos de uma estratégia capaz de medir o poder de viralização entre os usuários dentro deste ecossistema. Baseado na topologia de redes sociais somada às descobertas a partir dos dados coletados de grupos públicos de WhatsApp, nossa metodologia busca modelar como a informação de propaga no ambiente do WhatsApp e simular o efeito de viralização de mensagens tentando emular as limitações de *forward* e *broadcast* que ocorrem dentro desta rede. Para isso, utilizamos o modelo suscetível-infectado (SI) para realização de experimentos que medem a velocidade e alcance desse espalhamento de informação dentro das redes propostas, porém modificamos o modelo às necessidades do WhatsApp, adicionando um novo estágio intermediário, os expostos, propondo assim um modo Susceptível-Exposto-Infectado (SEI) mas adequado às peculiaridades desta rede.

Embora o modelo Suscetível-Infectado (SI) seja um dos modelos epidemiológicos mais simples, ele é bastante robusto em sua aplicação, e já foi utilizado para avaliar a disseminação de informação em redes sociais [Keeling and Eames, 2005]. Para analisar a propagação de informações em grupos do WhatsApp, utilizamos sua adaptação, o Suscetível-Exposto-Infectado (SEI) [Li and Zhen, 2005], considerando a desinformação como uma infecção que se espalha entre os usuários por meio da rede de grupos.

Nesse modelo, cada usuário é representado como um nó em uma rede de grupos, e os nós infectados podem disseminar a infecção para um grupo inteiro, expondo todos os seus participantes. O SEI modela três estágios: *Suscetível* (S), *Exposto* (E) e *Infectado* (I). No estágio *Suscetível*, os usuários ainda não tiveram contato com a infecção. No estágio *Exposto*, os usuários receberam a desinformação por meio de um dos grupos dos quais participam, mas ainda não a compartilharam. No estágio *Infectado*, os usuários que foram expostos à informação compartilham essa mensagem na rede em outros grupos.

O nosso modelo SEI ainda possui três parâmetros principais: *viralidade* (α), *exposição* (β) e um *limite de encaminhamento* (φ). A *viralidade* controla a taxa de usuários infectados, representando a probabilidade de um usuário infectado compartilhar o conteúdo com seus contatos. A *exposição* é a taxa na qual usuários expostos se tornam infectados, ou seja, a probabilidade de um usuário exposto se tornar um usuário infectado. O *limite de encaminhamento* é um parâmetro que restringe a propagação da infecção, simulando as limitações de encaminhamento de mensagens no WhatsApp. Ele determina o número máximo de grupos para os quais um usuário infectado pode enviar o conteúdo.

A simulação do modelo SEI é iniciada selecionando aleatoriamente um usuário como o nó inicial infectado. A cada iteração, os usuários expostos têm uma probabilidade α de compartilhar a mensagem maliciosa. Quando um nó infectado decide encaminhar, ele está sujeito ao limite de encaminhamento φ, o qual define o número máximo de grupos para os quais o conteúdo será enviado. Após o encaminhamento, os usuários nos grupos que receberam a mensagem ficam expostos. Em seguida, cada usuário exposto tem uma probabilidade β de se tornar um nó infectado e compartilhar o conteúdo. Essa iteração continua até que todos os usuários estejam infectados.

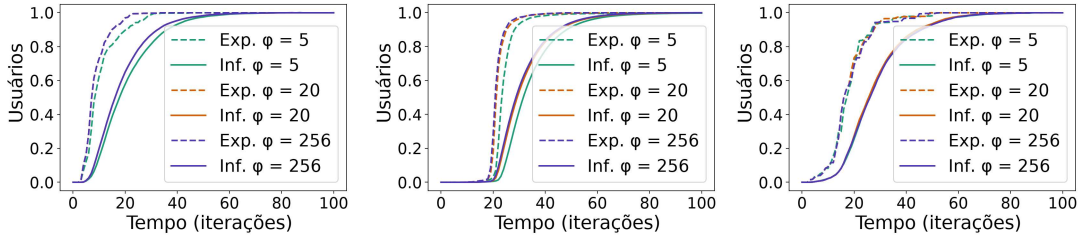
Ao adaptar o modelo SEI para o contexto do WhatsApp, consideramos os estágios de suscetibilidade, exposição e infecção, além de incorporar o limite de encaminhamento como uma restrição à propagação da informação. Essa abordagem permite avaliar a velocidade e o alcance da disseminação de informações nos grupos do WhatsApp, levando em consideração as características da plataforma, como o compartilhamento limitado e as restrições de encaminhamento. Através da manipulação dos parâmetros α, β e φ, podemos analisar os efeitos dessas restrições na viralidade das mensagens e entender como a informação se espalha dentro da rede do WhatsApp.

Desta maneira mais formal, considerando o período de incubação relativamente pequeno e a população constante, ou seja, sem nascimentos, mortes ou migrações, de tamanho N tem-se:

$$N = S(t) + E(t) + I(t) \tag{1}$$

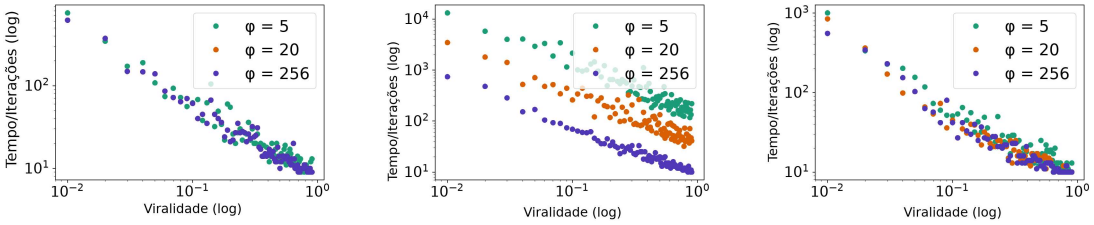
Levando-se em conta que a variação da população infectada é proporcional à população exposta, o sistema de equações diferenciais que descreve a dinâmica desta epidemia no modelo SEI é dado por:

$$\begin{cases} \dfrac{dS}{dt} = -\alpha SE & ,\alpha > 0 \\ \dfrac{dE}{dt} = \alpha SE - \beta E & ,\beta > 0 \\ \dfrac{dI}{dt} = \beta E \end{cases} \tag{2}$$



(a) Brasil (b) India (c) Indonésia

Figura 1.13. Simulações do modelo SEI variando a restrição de encaminhamento entre grupos (φ) com $\alpha = \beta = 0.1$.



(a) Brasil (b) Índia (c) Indonésia

Figura 1.14. Tempo para infectar todos os usuários da rede em simulações do modelo SEI variando a viralidade (α) de 0.001 até 1.0 .

Realizamos vários experimentos usando nosso modelo SEI para comparar a disseminação em diferentes cenários, aplicando restrições de transmissão e encaminhamento. Para essas simulações, consideramos apenas o maior componente conectado, uma vez que não seria possível alcançar nós isolados utilizando toda a estrutura da rede.

A Figura 1.13 mostra a fração de usuários infectados ao longo do tempo para todos os países quando o limite de encaminhamento (φ) é variado, ou seja, como as restrições implementadas pelo WhatsApp podem interferir na propagação. Consideramos o limite de encaminhamento para 5 grupos (cenário real), 20 grupos (limite anterior) e 256 grupos (limite atual para transmissão em massa). Observe que a taxa de usuários expostos na rede cresce muito rapidamente, independentemente dos limites de encaminhamento, sendo suficiente apenas 60 iterações para infectar toda a rede. Além disso, as limitações no encaminhamento diminuem ligeiramente a velocidade de propagação, mas não a interrompem completamente, especialmente para usuários expostos.

Também avaliamos o tempo necessário para que (des)informações com diferentes potencialidades virais infectem todos os usuários. A Figura 1.14 mostra o tempo necessário para infectar 100% dos usuários variando α de $0,001$ até $1,0$, com diferentes limites de encaminhamento. Observe que em situações de disseminação em massa (alto α), é difícil interromper a infecção devido às fortes conexões entre os grupos. No entanto, observe que os limites de encaminhamento e transmissão ajudam a retardar a propagação, principalmente em redes maiores, como na Índia. Em resumo, *os limites de encaminhamento e transmissão podem reduzir a velocidade de disseminação em uma ordem de magnitude para qualquer valor de viralidade α, entretanto não são capazes de bloquear que o conteúdo viralize pela rede*.

1.6. Abordagens para Detecção de Desinformação

Uma forma eficaz de detectar desinformação compartilhada em plataformas digitais é a checagem direta de fatos, normalmente realizada por jornalistas especializados e/ou agência de checagem. Uma tarefa de verificação de fatos (i.e., a avaliação da veracidade de uma notícia, mensagem ou afirmação) [Vlachos and Riedel, 2014] verifica a exatidão das informações comparando-as com uma ou mais fontes confiáveis. Exemplos de tais organizações incluem "Snopes.com"[38], "PolitiFact"[39], "FactCheck. org"[40], e "Aos fatos"[41], " é ou não é (G1)"[42], "Lupa"[43], "Boatos.org"[44] e "Projeto Comprova"[45], no Brasil.

No entanto, apesar da inegável importância neste contexto, a checagem de fatos é um processo demorado, pois geralmente requer uma análise detalhada para apoiar o veredicto [Vlachos and Riedel, 2014]. Consequentemente, a checagem tradicional de fatos não consegue acompanhar o enorme volume de informações que agora são geradas diariamente no ambiente online, que incluem plataformas digitais. Assim, estão emergindo vários estudos focados na verificação computacional de fatos [Atanasova et al., 2019], incluindo detecção automática de desinformação em diferentes cenários e plataformas [Conroy et al., 2015; Volkova et al., 2017]. Diante deste contexto, abordagens automáticas para a detecção de desinformação podem ser úteis para auxiliar, por exemplo, as agências de checagem de fatos na identificação de um conteúdo que necessite ser checado, sugerindo que o veredito final ainda dependa de um especialista. De forma geral, os esforços emergentes neste cenário podem ser divididos em dois grandes grupos: *(i)* estudos propõem soluções baseadas em técnicas de aprendizado de máquina; e *(ii)* esforços que exploram ferramentas ou sistemas online para suporte no monitoramento de (des)informação.

1.6.1. Soluções Baseadas em Aprendizado de Máquina

Essencialmente, notícias contendo desinformação são um viés de distorção nas informações manipuladas pelo editor/produtor do conteúdo [Shu et al., 2017]. Esforços anteriores sobre a teoria do viés de mídia [Gentzkow et al., 2015] mostram que o viés de distorção é geralmente modelado como um problema de classificação binária. Além disso, também há esforços relacionados que exploraram a detecção de desinformação (ou notícias falsas) como uma tarefa binária [Shu et al., 2017; Conroy et al., 2015; Wang, 2017; Volkova et al., 2017; Reis et al., 2019b]. Assim, com base nessas principais razões, no escopo deste capítulo, também definimos a detecção de desinformação como um problema de classificação binária em que a tarefa do classificador é distinguir notícias contendo desinformação das demais (e.g., notícias verdadeiras). Formalmente, o problema pode ser definido na seguinte forma:

[38] www.snopes.com
[39] www.politifact.com/
[40] www.factcheck.org/
[41] aosfatos.org
[42] g1.globo.com/e-ou-nao-e/
[43] piaui.folha.uol.com.br/lupa/
[44] www.boatos.org
[45] projetocomprova.com.br/

Definição 1.6.1 ***(Detecção de Desinformação.)*** *Dada uma notícia/mensagem não rotulada $a \in \mathscr{A}$, um modelo para detecção de desinformação atribui uma pontuação $S(a) \in [0,1]$ indicando até que ponto se acredita que a contenha desinformação. Por exemplo, se $S(a') > S(a)$, de acordo com o modelo é mais provável que a' contenha desinformação em comparação com a. Neste cenário, um limite τ pode ser definido de forma que a função de previsão $F : \mathscr{A} \rightarrow \{$desinformação, não contém desinformação$\}$ seja:*

$$F(a) = \begin{cases} \textit{desinformação} & \textit{if } S(a) > \tau, \\ \textit{não contém desinformação} & \textit{caso contrário.} \end{cases}$$

Assim, há vários esforços que propõem soluções baseadas em técnicas de aprendizado de máquina, como supervisionado [Conroy et al., 2015; Pratiwi et al., 2017; Wang, 2017; Volkova et al., 2017; Reis et al., 2019b], por reforço [Wang et al., 2020], ativo [Bhattacharjee et al., 2017] e profundo [Ruchansky et al., 2017; Zhang et al., 2018; Wang et al., 2018b; Kumar et al., 2020; Chen et al., 2023], e também, com base em estratégias específicas como blockchain [Paul et al., 2019]. Por exemplo, Pérez-Rosas et al. [Pérez-Rosas et al., 2017] conduzem um conjunto de experimentos de aprendizado para construir detectores precisos de notícias falsas usando conjuntos de recursos linguísticos. Da mesma forma, Volkova et al. [Volkova et al., 2017] constroem modelos linguísticos para classificar notícias suspeitas e confiáveis. De forma geral, a maioria desses esforços reduz o problema a uma tarefa de classificação, na qual as notícias são rotuladas como verdadeira/falsa e uma técnica de aprendizado de máquina é então usada para separar o conteúdo falso (ou desinformativo) dos demais com uso de um modelo "aprendido" a partir dos dados de treinamento. Especificamente, esses estudos comumente identificam padrões (ou atributos) recorrentes em notícias contendo desinformação depois que elas já foram disseminadas para propor novos recursos para treinar esses modelos a partir de dados específicos. Os principais atributos explorados na literatura para a proposição dessas abordagens são apresentados na seção a seguir.

1.6.1.1. Atributos para Detecção de Desinformação

A literatura é bastante ampla se considerarmos os esforços relacionados à credibilidade da informação, detecção de boatos, rumores e divulgação de notícias. Assim, conduzimos um levantamento sistemático desses esforços visando identificar os principais atributos propostos e explorados por trabalhos anteriores para detecção de desinformação. A Tabela 1.3 apresenta um resumo deste levantamento junto com algumas das técnicas utilizadas para extração desses atributos. Em alto nível, podemos categorizá-los da seguinte forma: *(i)* atributos extraídos do conteúdo da notícia (e.g., características do texto) [Gupta et al., 2014; Zhao et al., 2015; Wei and Wan, 2017; Volkova et al., 2017]; *(ii)* atributos da fonte da informação (e.g., confiabilidade e credibilidade) [Li et al., 2015]; e por fim *(iii)* atributos extraídas do ambiente, que geralmente envolve medidas de propagação do conteúdo dentro e fora das plataformas digitais [Ciampaglia et al., 2015]. É importante mencionar que, além de serem úteis para a proposição de abordagens tecnológicas nesta temática, esses atributos podem ser explorados para melhor entendimento da desinformação em diferentes contextos.

Tabela 1.3. Visão geral dos atributos para detecção de desinformação explorados em trabalhos anteriores.

Extraído do(a)...	Grupo de Atributos	Técnicas mais utilizadas/exemplos de atributos	Referências
Conteúdo	Atributos Sintáticos	Atributos em nível de sentença, indicadores de qualidade do texto (ex.: métricas de legibilidade), etc	[Conroy et al., 2015; Shu et al., 2017; Rubin et al., 2016; Ratkiewicz et al., 2011; Wei and Wan, 2017; Kwon et al., 2017]
	Atributos Lexicais	Atributos em nível de caracteres e palavras, incluindo número de palavras, pronomes, verbos, indicadores do uso de hashtags, pontuações, etc	[Castillo et al., 2011; Shu et al., 2017; Bhattacharjee et al., 2017; Gupta et al., 2014; Wei and Wan, 2017; Zhao et al., 2015; Kumar et al., 2016; Ribeiro et al., 2017; Ratkiewicz et al., 2011; Ahmed et al., 2017]
	Fundamentos Morais	Atributos ou medidas de fundamentos morais	[Volkova et al., 2017]
	Imagens e Vídeos	Propriedados associadas às imagens e aos vídeos (e.g., distribuições, indicadores de manipulação, etc)	[Jin et al., 2017]
	Atributos Psicolinguísticos	Sinais adicionais de linguagem persuasiva, como raiva, tristeza, etc. e indicadores de linguagem tendenciosa	[Volkova et al., 2017; Rubin et al., 2016; Vosoughi et al., 2018; Gupta et al., 2014; Kwon et al., 2017]
	Estrutura Semântica	*Word embeddings*, modelagem de tópicos (e.g., *Latent Dirichlet allocation (LDA)*), informações contextuais, medição de toxicidade do texto	[Friggeri et al., 2014; Conroy et al., 2015; Bhattacharjee et al., 2017; Rubin et al., 2016; Wei and Wan, 2017; Zhao et al., 2015; Ciampaglia et al., 2015; Wang, 2017]
	Subjetividade	Medidas de subjetividade e análise de sentimentos	[Volkova et al., 2017; Rubin et al., 2016; Ratkiewicz et al., 2011]
Fonte	Editor e Viés	Informações do produtor de conteúdo, indicadores de viés (e.g. político), polarização	[Ribeiro et al., 2017]
	Credibilidade e Confiabilidade	Estimativa da percepção do usuário sobre a credibilidade/confiabilidade da fonte	[Castillo et al., 2011; Shao et al., 2018; Shu et al., 2017]
Ambiente (Plataforma Digital e Web)	Engajamento (Interno e Externo)	Número de compartilhamentos do conteúdo, medidados (dentro e fora da plataforma digital), etc	[Castillo et al., 2011; Shu et al., 2017; Tacchini et al., 2017; Finn et al., 2014; Shao et al., 2016; Vosoughi et al., 2018; Friggeri et al., 2014; Gupta et al., 2014; Kumar et al., 2016]
	Estrutura da Rede	Conexões/redes de amizade, métricas de redes complexas	[Shao et al., 2018; Conroy et al., 2015; Shu et al., 2017; Volkova et al., 2017; Shao et al., 2016; Vosoughi et al., 2018; Castillo et al., 2011; Ratkiewicz et al., 2011; Friggeri et al., 2014; Gupta et al., 2014; Kumar et al., 2016; Tschiatschek et al., 2018]
	Padrões Temporais e Novidade	Séries temporais, medidas de propagação, métricas de novidade	[Castillo et al., 2011; Shao et al., 2018; Shu et al., 2017; Finn et al., 2014; Shao et al., 2016; Vosoughi et al., 2018; Friggeri et al., 2014; Kwon et al., 2017; Tschiatschek et al., 2018]
	Informação dos Usuários	Perfis e características dos usuários em diferentes níveis (e.g., individual, grupos, etc)	[Castillo et al., 2011; Shao et al., 2018; Shu et al., 2017; Tacchini et al., 2017; Ribeiro et al., 2017; Shao et al., 2016; Vosoughi et al., 2018; Ratkiewicz et al., 2011; Gupta et al., 2014; Kwon et al., 2017; Tschiatschek et al., 2018]

1.6.2. Sistemas para Monitoramento da (Des)Informação

Finalmente, surgiram vários sistemas com objetivo de monitorar o conteúdo disseminado em plataformas digitais, como contramedidas ao problema da desinformação nesses ambientes. Exemplos de tais sistemas incluem *(i)* "Hoaxy" [Shao et al., 2016], uma plataforma da Web para o rastreamento de notícias compartilhadas contendo desinformação; *(ii)* "Fake Tweet Buster" [Saez-Trumper, 2014], uma ferramenta da Web para identificar usuários que promovem desinformação no Twitter; e *(iii)* ELEIÇÕES SEM FAKE[46] no Brasil, nosso projeto, detalhado na seção a seguir, desenvolvido para trazer transparência na divulgação de conteúdo durante as eleições brasileiras em 2018, como um esforço para mitigar e evitar a disseminação de desinformação bem como prover transparência no espaço midiático.

1.7. Relato de Experiência: Projeto Eleições Sem Fake

Nesta seção apresentamos um relato de experiência sobre o projeto ELEIÇÕES SEM FAKE, proposto e coordenado pelo professor Fabrício Benevenuto do Departamento de Ciência da Computação (DCC) da Universidade Federal de Minas Gerais (UFMG), em colabora-

[46] www.eleicoesemfake.dcc.ufmg.br

ção com estudantes e professores de diversas instituições brasileiras, incluindo a Universidade Federal de Viçosa (UFV), a Universidade Federal de Mato Grosso do Sul (UFMS), dentre outras. Este projeto surgiu em 2018 como uma iniciativa para mitigar o problema da desinformação em plataformas digitais durante o período eleitoral daquele ano. As propriedades de rápida difusão e ampla propagação de informações nesses ambientes podem ser abusadas para fins de propaganda não solicitada, interrupção de comunicação legítima ou mesmo para manipulação de opinião. Assim, o objetivo do projeto foi trazer transparência para as campanhas políticas realizadas em plataformas digitais online, explorando soluções tecnológicas capazes de expor dados de campanhas de desinformação dentro deste contexto. Através da transparência, jornalistas, checadores de fatos, legisladores e consultores jurídicos puderam (e podem) atuar e permitir ações do poder público, caso sejam constatadas irregularidades.

Por exemplo, durante as eleições brasileiras de 2018, foram desenvolvidos sistemas que coletam, processam e analisam dados em larga escala de plataformas sociais tradicionais, como Twitter e Facebook, e especialmente do WhatsApp através do Monitor de WhatsApp [Melo et al., 2019a]. Este sistema foi extensivamente utilizado por jornalistas e equipes de checagens de fatos durante o processo eleitoral supracitado, facilitando o acesso aos dados da plataforma e permitindo uma navegação por data pelos conteúdos mais populares compartilhados por dia. Com isso, os usuários, conseguem apontar padrões e movimentos que emergem dentro do WhatsApp de uma forma que seria impossível sem o sistema, fornecendo um valioso suporte para a tarefa de checagem de fatos e reduzindo o esforço necessário para encontrar as notícias e desinformações que vão sendo viralizadas na rede.

Acreditamos que esse relato de experiência sobre o projeto ELEIÇÕES SEM FAKE seja extremamente relevante para quem deseja entender e combater a desinformação em plataformas digitais, principalmente durante períodos eleitorais. O relato mostra como a tecnologia pode ser usada para identificar e expor campanhas de desinformação, explorando de forma inteligente um grande volume de dados que é diariamente disseminado nesses ambientes. Além disso, o projeto se mostrou eficaz ao fornecer um suporte valioso para jornalistas e equipes de checagem de fatos, reduzindo o esforço necessário para encontrar notícias e desinformações viralizadas nestes ambientes e também no desenvolvimento de pesquisas acadêmicas que recorrem aos dados disponibilizados por meio dos diversos sistemas desenvolvidos. Logo, descrever experiências relacionadas ao projeto em questão pode ser útil para pessoas interessadas em desenvolver soluções tecnológicas para combater a desinformação, para jornalistas e profissionais da comunicação que desejam aprimorar sua cobertura de notícias em períodos eleitorais, bem como para pesquisadores de diferentes áreas do conhecimento interessados no estudo do fenômeno. Especificamente nas seções a seguir nos atemos a apresentar em detalhes dois dos principais sistemas que compõem o projeto, que chamamos: *(i)* MONITOR DE WHATSAPP (ou "*WhatsApp Monitor*") e o *(ii)* "Monitor de Anúncios do Facebook", referenciado como ADCOLLECTOR.

1.7.1. Monitor de WhatsApp

Infelizmente, as plataformas plataformas se tornaram ambientes propícios para a propagação de campanhas de desinformação, especialmente no contexto político. O WhatsApp,

por exemplo, com seus mais de 2 bilhões de usuários, desempenha um papel central nesse cenário, tendo sido palco de eventos de desinformação em diferentes partes do mundo. Durante as eleições presidenciais de 2018 no Brasil, por exemplo, houve uma disseminação massiva de desinformação por meio da plataforma, despertando grande preocupação sobre como o WhatsApp opera [Tardaguila et al., 2018].

Um dos principais desafios para combater a disseminação de desinformação no WhatsApp é a dificuldade em analisar o conteúdo compartilhado dentro da plataforma. Embora uma grande parte das conversas ocorram em grupos públicos com até 256 membros, o acesso restrito e a criptografia ponta-a-ponta dificultam a exploração desse conteúdo viral. Entretanto, ao contrário de outros sistemas mais tradicionais, pesquisadores e jornalistas não possuem uma ferramenta abrangente para analisar as mensagens mais populares que circulam nessa plataforma, o que torna o combate à desinformação neste cenário um desafio ainda maior. Nesse contexto, o monitor de WhatsApp, que depois foi replicado também para dados do Telegram Júnior et al. [2022], permite aos usuários explorar o conteúdo mais compartilhado em grupos públicos de aplicativos de mensageria. Com a implementação destas ferramentas, oferecemos uma forma eficaz de combater a disseminação de desinformação, auxiliando pesquisadores, jornalistas e demais interessados a analisar o conteúdo que viraliza nessa plataforma de mensagens instantâneas.

De forma resumida, nossa ferramenta monitora várias categorias de mensagens, como imagens, vídeos, áudio e texto que foram postadas em um conjunto de centenas de grupos públicos políticos do WhatsApp e exibe os conteúdos mais compartilhados por dia numa interface online. Ela já foi usada para monitorar conteúdo durante as eleições gerais brasileiras de 2018, eleições de 2019 na Índia e na Indonésia, para acompanhar notícias e outras (des)informações sobre a pandemia COVID-19 durante 2020 e 2021 e também acompanhar o conteúdo compartilhado no WhatsApp durante as eleições presidenciais de 2022. Este sistema é um dos principais esforços atualmente para estimar a disseminação de desinformação no WhatsApp e ajudar nos esforços de verificação de fatos dentro deste cenário [Melo et al., 2019a]. A arquitetura do sistema é responsável por agrupar, coletar, processar e ranquear o conteúdo explorado do WhatsApp e condensa-lo numa interface online, acessível em um navegador de Internet através de login e senha para ser possível o usuário final navegar entre os dados coletados e fazer suas próprias análises com a ajuda do sistema. O principal objetivo do MONITOR DE WHATSAPP é propiciar um sistema capaz de informar e antecipar aos comunicadores sobre o tipo de informação compartilhada no WhatsApp. A seguir, apresentamos maiores detalhes relativos ao sistema. Em seguida, discutimos os impactos da utilização da ferramenta criada e, por último, relacionamos nossas considerações finais acerca do MONITOR DE WHATSAPP.

1.7.1.1. Metodologia

O nosso sistema Web MONITOR DE WHATSAPP foi idealizado originalmente em 2018 numa versão preliminar publicada por Resende et al. [Resende et al., 2018]. Uma versão funcional da ferramenta foi desenvolvida e disponibilizada no ano seguinte [Melo et al., 2019a]. Depois, a proposta foi também replicada no Telegram [Júnior et al., 2022]. Atualmente, o MONITOR DE WHATSAPP conta com uma versão online com atualizações diárias de conteúdo com todas as funcionalidades de *ranking* e agrupamento aprimora-

das. Sua estratégia de coleta segue os mesmos passos apresentados anteriormente neste capítulo, na seção de coleta de dados de aplicativos de mensagens instantâneas. Portanto, explicaremos de forma bastante reduzida esta etapa de como ele funciona novamente aqui, enquanto focaremos mais em outros aspectos como a interface implementada e o impacto de sua utilização.

O MONITOR DE WHATSAPP usa de dados coletados de grupos públicos do WhatsApp selecionados que discutem tópicos políticos. Esses grupos públicos são operados tanto por indivíduos afiliados a partidos políticos, líderes comunitários locais ou usuários comuns com interesse no tema, e podem ser acessados livremente por qualquer pessoa com um link de convite: uma URL do WhatsApp no padrão `chat.whatsapp.com/<groupID>` compartilhada em plataformas sociais para qualquer pessoa que deseje entrar no grupo.

A primeira etapa do trabalho é, portanto, selecionar os grupos de interesse que serão monitorados pelo sistema. Enquanto os dados de grupos da Índia e Indonésia foram adquiridas em parceria nossa com pesquisadores do MIT [Garimella et al., 2018], para o Brasil seguimos a coleta de grupos públicos de política no WhatsApp. Para o levantamento dos grupos públicos, utilizamos uma lista de palavras-chaves, junto com a URL de convite do WhatsApp, numa pesquisa em plataformas sociais (e.g., Twitter e Facebook) e em ferramentas de busca (e.g., Google) para encontrar grupos relevantes para o monitoramento. Após listar os resultados, configuramos três celulares com contas válidas do WhatsApp para participar de cada grupo. Após entrar, expandimos nossa coleção com outros grupos compartilhados dentro dos chats de conversa do próprio WhatsApp, resultando em uma coleção de mais de mil grupos selecionados.

Posteriormente, com toda configuração preparada, iniciamos a extração de dados, coletando e processando todos as mensagens dos *chats* no servidor conforme a arquitetura de coleta proposta neste capítulo. O processo de coleta é feito diariamente, alimentando o banco de dados com conteúdo por dia com todos os textos e mídias relacionadas. Cada conta utilizada é membro propriamente dito de um conjunto específico de grupos que recebe, tal como qualquer outro usuário de WhatsApp, todas suas mensagens em seu celular. Para coleta propriamente dita, não existe uma API oficial do WhatsApp para este tipo de aplicação de coleta de mensagens. Desta forma, utilizamos de uma estratégia de raspagem de dados através do WhatsApp Web com auxílio de uma biblioteca WebWhatsAPI[47]. Essa ferramenta utiliza a versão de navegador do WhatsApp para fazer um *parsing* da página Web e coletar de um usuário os grupos e todo o conteúdo recebido em cada um. Esse conteúdo é processado e organizado no banco de dados de forma que, quando um usuário do MONITOR DE WHATSAPP navega no *website*, o sistema acessa esse banco e exibe um *dashboard* com as mídias e textos referentes ao período visualizado.

Uma das etapas mais importantes na elaboração e criação de um monitoramento de dados de plataformas de mensagens instantâneas é o agrupamento de conteúdo semelhante, o que permite rastrear e contar quantas vezes um determinado conteúdo, como imagens, vídeos ou áudios, foi compartilhado. Diferentemente de outras plataformas, onde as informações sobre curtidas e compartilhamentos já estão disponíveis nos metadados, no WhatsApp cada mensagem é postada de forma isolada. Portanto, é necessário

[47] `https://github.com/mukulhase/WebWhatsapp-Wrapper`

Figura 1.15. *Screenshot* da interface principal do MONITOR DE WHATSAPP com exibição do conteúdo mais popular de um determinado dia.

realizar o rastreamento, agrupamento e contagem dos dados para determinar a popularidade de cada conteúdo compartilhado.

Para rastrear os compartilhamentos de um único conteúdo e contar sua popularidade, utilizamos o *PerceptualHash* (*pHash*) para gerar uma impressão digital única para cada imagem [Du et al., 2020]. Para áudios e vídeos, calculamos um *hash* a partir do *checksum* (MD5) de cada arquivo. Com esses *hashes*, podemos agrupar as postagens que contêm o mesmo conteúdo e calcular informações sobre ele, como sua popularidade, quantos grupos o compartilharam e quantos usuários diferentes o enviaram. Além disso, usamos o índice Jaccard para comparar mensagens de texto e agrupá-las. Essa contabilização é realizada diariamente, e os usuários podem acessar o *ranking* de popularidade por meio da interface Web do MONITOR DE WHATSAPP.

Questões Éticas. O MONITOR DE WHATSAPP reúne uma quantidade considerável de dados de muitos grupos e usuários do WhatsApp. Para garantir a privacidade dos usuários, não compartilhamos ou divulgamos quaisquer informações de identificação pessoal, como números de telefone celular ou nome do usuário. Outro material sensível que pode estar presente em nosso conjunto de dados são imagens que retratam violência explícita ou conteúdo adulto. Como pode ser prejudicial para os usuários que navegam em nosso sistema, utilizamos um filtro de conteúdo adulto para detectar esse tipo de conteúdo e rotulá-lo no sistema. Para evitar o uso indevido, mesmo de informações agregadas, também limitamos o acesso do nosso sistema a um número restrito de jornalistas e pesquisadores, por meio de uma conta de login e senha. Além disso, eles também são informados sobre as limitações dos dados e o potencial viés presente em nosso sistema. Como usamos apenas grupos do WhatsApp disponíveis publicamente, assim como todos os demais membros, nossa coleta de dados não viola os termos de serviço do WhatsApp.

1.7.1.2. Interface e Uso do sistema

Nós fornecemos um sistema online no qual os usuários podem supervisionar diariamente as tendências compartilhadas nos grupos públicos do WhatsApp. Nosso sistema exibe informações sobre as mídias de texto (apenas os com mais de 140 caracteres), áudio,

Figura 1.16. Painel do MONITOR DE WHATSAPP em que os usuários podem selecionar a data ou escolher um período.

imagem e vídeo dos grupos relacionadas a tópicos políticos e de notícias monitorados diariamente. Em seguida, ranqueamos cada tipo de mídia entre as mais populares. A Figura 1.15 mostra uma captura de tela de como o conteúdo é exibido na interface do MONITOR DE WHATSAPP assim que o usuário faz login no sistema, mostrando, em ordem, os conteúdos com mais compartilhamentos. Vale mencionar que o usuário pode escolher também entre português ou inglês como idioma da interface no canto superior esquerdo da tela.

Atualmente, nosso sistema funciona para três instâncias distintas: uma indiana, uma versão indonésia e uma brasileira[48]. Depois que uma instância é escolhida e o usuário faz login, ele é levado a um painel de controle onde pode navegar entre as datas e observar o conteúdo mais compartilhado, conforme mostrado na Figura 1.16. O sistema também permite que os usuários selecionem o período que desejam, como, por exemplo, um dia, uma semana ou mês inteiro de dados. Depois de escolher uma data de início e término para a pesquisa, o sistema recupera e relata o conteúdo mais popular para toda a data selecionada. Isso permite a jornalistas e pesquisadores investigarem um período específico ou mesmo eventos que duram mais de um dia, combinando milhares de mensagens em uma interface resumida e ranqueada, na qual podem surgir alguns padrões de publicação e conteúdo que, sem o sistema, seria difícil perceber. Isso permite que os jornalistas tenham uma melhor ideia sobre o conteúdo crítico compartilhado no WhatsApp que pode valer a pena ser verificado.

No MONITOR DE WHATSAPP, quatro tipos de conteúdo são identificados e consolidados no banco: imagens, vídeos, mensagens de áudio e mensagens de texto (apenas as com mais de 140 caracteres). Nosso sistema exibe diariamente o conteúdo dividido por cada tipo de mídia e os mostra cada uma ranqueada por total de compartilhamentos. Isso permite que os jornalistas tenham diariamente uma ideia sobre o conteúdo crítico compartilhado no WhatsApp que pode valer a pena ser verificado.

Para dar mais detalhes sobre cada conteúdo compartilhado no WhatsApp, ao clicar em um objeto no painel, disponibilizamos informações compiladas de compartilhamentos entre os grupos monitorados para cada conteúdo selecionado. Um usuário tem à sua dis-

[48]A versão indiana conta apenas com dados estáticos de 2019 durante as eleições gerais indianas, enquanto o Brasil e a Indonésia continuam sendo atualizadas diariamente.

posição o número total de compartilhamento, em quantos grupos esse conteúdo apareceu e quantos usuários únicos postaram sobre aquele conteúdo. Além disso, existe uma funcionalidade de busca com botão "Na Web" para outras fontes. Clicando em uma imagem, é possível verificar o conteúdo com botão, que usa da busca reversa de imagens do Google para rastrear fontes externas onde aquela imagem foi compartilhada. Curiosamente, o próprio aplicativo WhatsApp implementou uma funcionalidade muito semelhante, chamada "Pesquisar na Web" em 2020, em que usuários podem pesquisar por conteúdo viral na Web[49].

1.7.1.3. Discussões e Impacto Científico-Social do MONITOR DE WHATSAPP

Desde as eleições brasileiras de 2018 até julho de 2021, demos acesso ao sistema a mais de 300 usuários, incluindo jornalistas, pesquisadores e agências de checagem de fatos que mencionaram explicitamente nosso sistema como fonte de dados durante as checagens. Adicionalmente, dezenas de notícias fizeram referência ao nosso sistema ou usaram nossos dados durante as eleições brasileiras e durante a pandemia de COVID-19 para entender melhor as discussões que ocorrem dentro do WhatsApp. Mais especificamente, matérias da BBC, The Guardian, El Pais, The Intercept, O Globo, Estadão, Folha, Uol, entre diversas outras que utilizam do sistema para investigar o WhatsApp e produzir a reportagem, conforme mais detalhado em [Melo et al., 2021].

Ademais, o MONITOR DE WHATSAPP foi capaz de agrupar conteúdo a partir de um grande volume de dados e ranqueá-lo diariamente até o uso de muitos jornalistas e agências. Notavelmente, nosso sistema é referenciado como parceiro do Comprova[50], um projeto jornalístico da *First Draft* com foco na verificação de conteúdo publicado na Web durante a eleição presidencial brasileira de 2018 e pela agência de checagem de fatos Lupa[51] de jornalistas do grupo Folha. Em 2018, a UFMG foi parceira do TSE através do nosso sistema para conter a desinformação nas eleições[52] e, em 2020, tivemos uma parceria com o Ministério Público de Minas Gerais (MPMG), como um dos projetos do Programa de Capacidades Analíticas, que visa prover mais transparências sobre dados públicos online[53].

Além dos impactos e colaborações mencionados, diversos estudos, feitos fora do nosso grupo de pesquisa da UFMG, também utilizaram o MONITOR DE WHATSAPP como metodologia e fonte de dados para avançarem as pesquisas nas áreas do WhatsApp e também em torno da temática de desinformação no país. O sistema auxilia esses pesquisadores a desenvolverem seus trabalhos, provendo transparência e facilidade para navegar em períodos passados do WhatsApp, como por exemplos os trabalhos de [Recuero et al., 2021] e [da Silva, 2021], entre diversos outros que tiveram o sistema como fonte de dados para sua análise. Müzell [Müzell, 2020], por exemplo, em sua dissertação de mestrado

[49] https://blog.whatsapp.com/search-the-web/?lang=en

[50] https://projetocomprova.com.br/partner/monitor-de-whatsapp-ufmg/

[51] https://piaui.folha.uol.com.br/lupa/tag/ufmg/

[52] https://www.tse.jus.br/imprensa/noticias-tse/2018/Outubro/tse-estuda-possibilidade-de-firmar-parceria-com-universidade-para-inibir-fake-news-no-whatsapp

[53] https://www.mpmg.mp.br/comunicacao/noticias/mpmg-inicia-trabalhos-de-convenio-com-ufmg-para-ampliar-capacidade-de-analise-de-dados.htm

se baseou nos dados do sistema durante as eleições de 2018, identificando estratégias e padrões sobre como a campanha política ocorreu no WhatsApp e como elas interferiram na eleição. Almeida et al. [Almeida et al., 2019], para saber quais conteúdos foram veiculados nas eleições de 2018, também utilizou o MONITOR DE WHATSAPP para explorar características próprias da circulação dessas plataformas, considerando o espaço privado de troca de informações. Soares et al. [Soares et al., 2021], com auxílio de nossa ferramenta, também estudaram a desinformação sobre COVID-19 no WhatsApp, observando como a pandemia enquadrada como debate político. Em outra direção, o estudo de Tomás et al. [Tomás et al., 2020] investigou as notícias falsas contra as universidades públicas no Brasil usando do MONITOR DE WHATSAPP.

O MONITOR DE WHATSAPP está online desde 2018, sendo atualizado diariamente com dados de mais de 900 grupos públicos sobre política na referida plataforma. Nosso sistema é utilizado como fonte de dados por vários pesquisadores e jornalistas, inclusive como fonte de três agências de checagem de fatos. A arquitetura do sistema coleta, processa, ranqueia e exibe todo conteúdo dos grupos de WhatsApp num sistema Web, hospedado no servidor do nosso grupo de pesquisa da UFMG e é acessível por navegador apenas através de usuário e senha. Nossa metodologia não só apresenta um modo inovador de coletar dados do WhatsApp [Resende et al., 2019], como também se mostrou eficaz na ajuda ao combate de desinformação. Com a transparência de acesso aos dados do WhatsApp e facilidade de uso com uma navegação por data pelos conteúdos mais populares compartilhados por dia, os usuários conseguem apontar padrões e movimentos que emergem dentro da plataforma de uma forma que seria impossível sem o sistema. Com isso, damos um valioso suporte para a tarefa de checagem de fatos, reduzindo o esforço necessário para encontrar as notícias e desinformações que vão sendo viralizadas na plataforma. Devido à natureza fechada e efêmera do WhatsApp, nosso sistema funciona como uma espécie de registro histórico dos eventos ocorridos na plataforma, uma vez que esses dados dificilmente seriam acessíveis de outra forma, dado que a empresa não armazena o conteúdo por muito tempo e mesmo usuários podem não ter mais registro das mensagens enviadas.

O sistema desenvolvido aqui pode ainda ser expandido futuramente para dar ainda mais suporte a cobertura dos eventos no WhatsApp durante próximos períodos eleitorais. Existe espaço para melhoras tanto na infraestrutura utilizada para coleta, com mais espaço no servidor da universidade e mais celulares para gerenciar contas de WhatsApp, como também o número de grupos monitorados. Pretendemos também integrar novas funcionalidades de ordenação e busca no sistema. Para fornecer novas formas de ranqueamento, estudaremos a implementação de métodos de aprendizado para detecção automática de desinformação, agregando uma pontuação (ou *score*) que indique a probabilidade do conteúdo ser conter desinformação. Esperamos que tal abordagem possa auxiliar ainda mais jornalistas encontrar conteúdos mais relevantes a serem checados. Por fim, também planejamos implementar a geração de relatórios periódicos referentes a datas específicas para facilitar a interpretação dos dados no sistema e fornecer informações mesmo para aqueles sem um cadastro no sistema.

1.7.2. Monitor de Anúncios do Facebook

Desde as eleições presidenciais dos Estados Unidos em 2016, o conteúdo patrocinado contendo Propaganda Eleitoral se tornou uma forma eficaz de campanha política. No entanto, essa eleição foi marcada pelo abuso da publicidade direcionada em Redes Sociais Online. Preocupados com o risco do mesmo tipo de abuso ocorrer nas eleições brasileiras de 2018, nós projetamos e implementamos uma abordagem computacional para monitorar anúncios políticos em plataformas de redes sociais.

Primeiramente, devido ao abuso da publicidade direcionada no Facebook (atual Meta) durante as eleições presidenciais dos Estados Unidos em 2016, escolhemos a plataforma do Facebook para validar nossa abordagem. Para isso, nós adaptamos inicialmente um *plug-in* do navegador para coletar anúncios da linha do tempo de voluntários que usam o Facebook. Nós conseguimos a adesão de mais de 2000 voluntários a ajudar em nosso projeto e instalar nossa ferramenta. Em seguida, utilizamos uma Rede Neural Convolucional (CNN) para detectar anúncios políticos no Facebook usando incorporação de palavras. Para avaliar nossa abordagem, rotulamos manualmente uma coleção de dados com 10 mil anúncios como políticos ou não políticos e, em seguida, realizamos uma avaliação detalhada da abordagem proposta para identificar anúncios políticos, comparando-a com métodos clássicos de aprendizado de máquina supervisionado.

1.7.2.1. Metodologia

Inicialmente, nosso principal objetivo era coletar anúncios diretamente da linha do tempo do usuário. Portanto, precisamos construir uma extensão de navegador cujo objetivo é coletar anúncios enquanto o usuário está navegando em seu perfil do Facebook. Além disso, nós coletamos anúncios da Biblioteca de Anúncios Políticos da Biblioteca de Anúncios do Facebook. Apesar das características históricas desses anúncios, eles contêm anúncios reais feitos por anunciantes políticos. Assim, a extensão do navegador e o coletor da Biblioteca de Anúncios pertencem ao módulo principal da arquitetura proposta chamado `AdCollector`.

A extensão de navegador consegue extrair as informações diretamente do HTML do anúncio. Nós extraímos informações como o nome do anunciante, identificador do anunciante, legendas do anúncio, imagens do anúncio, URLs e principalmente as informações fornecidas pelo Facebook sobre “Por que estou vendo isso?”. Ver Figura 1.17.

Figura 1.17. Exemplo de um “Por que estou vendo isso?”

Cada anúncio no Facebook inclui um botão e, quando os usuários clicam nele, eles recebem uma explicação sobre por que estão vendo aquele anúncio em particular. As explicações de anúncios geralmente têm duas partes: a primeira parte apresenta uma razão pela qual um usuário recebeu um anúncio. A redação geralmente indica o tipo de atributo que está sendo apresentado. A segunda parte lista os possíveis atributos que podem ter sido usados pelo anunciante para segmentar um anúncio. Geralmente inclui um conjunto limitado de informações, como idade, gênero e localização, usadas como opções de segmentação.

Página de Preferências de Anúncios. A página de preferências de anúncios é uma página personalizada que fornece aos usuários informações sobre vários atributos que influenciam os anúncios que eles veem, além de oferecer a eles um certo nível de controle sobre esses anúncios. Primeiro, os usuários podem ver os interesses que o Facebook inferiu sobre eles, como se eles estão interessados em coisas como "Jogos de Vídeo", "Pizza" ou até mesmo "Homossexualidade". Os interesses também são acompanhados por uma descrição de como eles foram inferidos. Essas descrições podem ser afirmações como "Você tem essa preferência porque clicou em um anúncio relacionado a *Interesse*" ou "Você tem essa preferência porque curtiu uma página relacionada a *Interesse*". Como mostrado em [Andreou et al., 2018], a grande maioria dessas explicações é vaga e não fornece muitos detalhes.

Nós utilizamos o ADCOLLECTOR para monitorar anúncios de 14 de março de 2018 a 28 de outubro de 2018. Esse período abrange o período eleitoral das elições, incluindo os dois turnos. No geral, mais de 2.000 usuários se voluntariaram para instalar nossa extensão de navegador e compartilhar os anúncios que receberam enquanto navegavam no Facebook com nosso plugin. Observamos, no entanto, que muitos usuários apenas instalaram a extensão do navegador, mas não a utilizaram. No entanto, um total de 715 usuários utilizaram ativamente nossa ferramenta nesse período.

A Figura 1.18 mostra o número de usuários ativos por dia. Nós consideramos um usuário ativo em um dia se ele recebeu pelo menos um anúncio do Facebook. O número de usuários diários aumentou rapidamente quando vários veículos de mídia publicaram artigos sobre nosso sistema e permaneceu relativamente estável posteriormente. A diminuição repentina de usuários ativos em meados de junho pode ser atribuída a uma mudança que o Facebook fez na forma como rotula os anúncios, resultando na perda de anúncios por alguns dias. Também observamos que a atividade dos usuários nos fins de semana diminui, o que pode indicar que alguns usuários instalaram nosso *plugin* em seus computadores no trabalho.

Dos 715 usuários, 682 são do Brasil. Nós inferimos essa informação de segmentação analisando os dados das explicações do "*Por que estou vendo isso?*" que foram coletadas pela nossa extensão. Coletamos um total de **239k anúncios únicos** enviados por **40k anunciantes**. Cada anúncio é identificado por um *id* único fornecido pelo Facebook. Dos 239k anúncios únicos, 166k foram enviados durante o período pré-eleitoral (março de 2018 a 15 de agosto de 2018) e 74k foram enviados durante o período eleitoral (16 de agosto de 2018 a 28 de outubro de 2018). Para cada anúncio, temos informações sobre o anunciante, o texto do anúncio, o texto no aviso político do anúncio, a imagem (quando disponível) e a URL de destino. Nos referimos a esse conjunto de dados como o

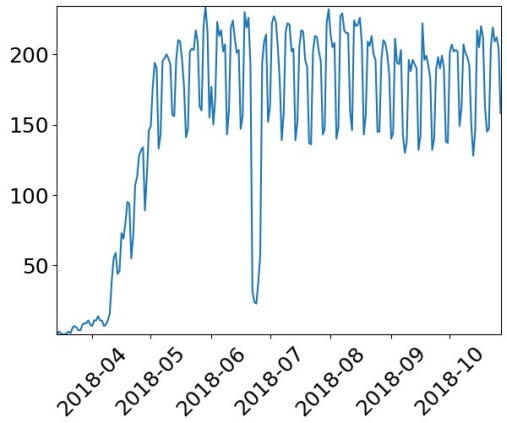


Figura 1.18. Número de usuários ativos diariamente.

ADCOLLECTORDATA.

Os anúncios políticos oficiais são diferentes dos anúncios regulares. Além da *tag* "Patrocinado", o aviso também contém a *tag* "Propaganda Eleitoral". Por causa disso, nossa extensão não coletou esses anúncios, ou seja, o conjunto de dados coletado com nossa extensão de navegador não contém nenhuma propaganda eleitoral que foi explicitamente informada ao Facebook pelos anunciantes.

Nós utilizamos o modelo CNN treinado com dados rotulados por três rotuladores independentes e atribuímos a cada anúncio uma probabilidade de ser político. Na prática, há um número significativamente maior de anúncios não políticos do que anúncios políticos (ou seja, temos um cenário de conjunto de dados desbalanceado). Para limitar o número de falsos positivos (ou seja, anúncios classificados erroneamente como políticos pelo nosso modelo, mas que na verdade são não políticos), escolhemos um limiar para declarar um anúncio como político que corresponde a uma taxa de 1% de falsos positivos (em vez de escolher o limiar típico de 0,5 para a probabilidade de ser político, que corresponde a uma taxa de falsos positivos de 8% e taxa de verdadeiros positivos de 96%). Utilizando um limiar de 0,97, nosso modelo CNN classifica 1.133 anúncios como políticos dos 38.110 anúncios testados - $2,9\%$ dos anúncios são políticos. Os 1.133 anúncios foram postados por 739 anunciantes.

Na Figura 1.19 é apresentado um anúncio identificado dentro da nossa base de dados do ADCOLLECTOR utilizando o nosso classificador. Neste anúncio, um candidato elogia o outro ressaltando suas qualidades, e no criativo do anúncio é apresentado o número do candidato. É importante salientar que as informações do CNPJ foram colocados no corpo do anúncio, porém não havia o rótulo oficial da Plataforma Meta indicando que era uma propaganda política. Desta forma, ao término desta campanha este anúncio não ficará disponível na biblioteca de anúncio da Plataforma Meta.

1.8. Desafios e Oportunidades de Pesquisa

Combater a desinformação é uma típica luta adversária. A cada eleição, por exemplo, as campanhas de desinformação exploram novas formas de manipular a opinião e novos mecanismos de defesa são criados visando ao menos mitigar as campanhas de desinformação. Aqui, apresentamos alguns desafios e oportunidades de pesquisa nesta temática.

Figura 1.19. Exemplo de anúncio de propaganda detectado pelo nosso classificador.

Compreensão do Ecossistema de (Des)Informação. As plataformas digitais estão passando por mudanças constantes que impactam diretamente a forma como as informações são disseminadas dentro desses ambientes. O WhatsApp, por exemplo, recentemente, implementou o recurso de comunidades, que potencializa o disparo massivo de mensagens dentro da plataforma. Outro desafio é identificar padrões de desinformação que permanecem inalterados em diferentes contextos e ambientes de disseminação, por exemplo, ao longo do tempo. Logo, é fundamental que sejam investigados os impactos ocasionados por essas mudanças no ecossistema de (des)informação considerando as diferentes plataformas e contextos (e.g., saúde, política, etc) a fim de que seja possível compreendê-los para criação de mecanismos de combate e/ou mitigação de seus efeitos, quando aplicável. Além disso, em vários casos, as plataformas digitais são utilizadas apenas como veículos (ou vitrine) para disseminação de conteúdos que na verdade são produzidos por *websites* externos dedicados exclusivamente à produção e disseminação de desinformação [Couto et al., 2022a,b]. Nesta direção, acreditamos que ainda há espaço de pesquisa para uma compreensão mais aprofundada desse ecossistema, potencialmente orquestrado e financiado, bem como para a proposição de abordagens efetivas neste contexto.

Propagandas Políticas. Embora as plataformas tenham criado mecanismos de transparência no contexto de propagandas políticas eles ainda são bastante limitados. Recentemente, um grupo de pesquisadores conseguiram impulsionar conteúdos 100% falsos nas eleições de 2022 na Plataforma de Anúncios da Meta, onde o anúncio afirmava que a data da eleição havia mudado[54]. Além disso, o anúncio foi pago com moeda estrangeira e criado por um usuário que estava em Londres. Portanto, ainda existe espaço para pesquisas e soluções tecnológicas com objetivo de mitigar o uso de plataformas de impulsionamentos dentro das diferentes plataformas digitais com a intenção de prejudicar ou tumultuar o processo eleitoral. Acreditamos que a regulação seja importante, no entanto isso não elimina a necessidade de proposição de soluções tecnológicas como mecanismos de apoio em diferentes contextos.

[54] https://www.globalwitness.org/en/campaigns/digital-threats/facebook-fails-tackle-election-disinformation-ads-ahead-tense-brazilian-election/

Detecção Automática de Desinformação. A desinformação não é mais disseminada exclusivamente em formato texto no ambiente online. Atualmente, a desinformação é propagada nas plataformas digitais por meio de imagens, vídeos (e.g., *deep fake*, memes, stickers, etc. Logo, é preciso que as abordagens automáticas sejam adequadas e/ou desenvolvidas para a detecção de desinformação em diferentes tipos de mídias. Além disso, acreditamos que ainda há espaço para investigação de estratégias mais sofisticadas, envolvendo aprendizado por transferência, federado, etc, ou ainda técnicas estado-da-arte de representação de conteúdo (e.g., *embeddings*) que ainda são pouco explorados neste contexto.

Explicabilidade de Abordagens Tecnológicas. Existem alguns estudos que visam investigar a explicabilidade de resultados iniciais promissores da detecção computacional de desinformação em plataformas digitais, ou seja, por que uma determinada notícia é classificada como desinformação (ou não) [Shu et al., 2019; Yang et al., 2019; Cui et al., 2019; Reis et al., 2019a; Lu and Li, 2020]. No entanto, este é um campo que ainda carece de esforços, considerando principalmente a multidisciplinaridade envolta na temática.

Investigação de Plataformas Emergentes. Por fim, existem várias ferramentas emergentes baseadas em Inteligência Artificial (IA) que, devido ao acesso massivo e a facilidade no uso sem critérios bem definidos apresentam potencial para serem exploradas como recurso para a propagação de desinformação. Como exemplos podemos citar a ChatGPT[55], e ferramentas *Text-to-Image*, como a DALL-E[56], Midjourney[57], e o *Stable Diffusion*[58] que recentemente foram exploradas para criação de informação inverídica que circulou em plataformas digitais[59]. Logo, neste contexto é fundamental entendermos como essas ferramentas emergentes estão/podem ser utilizadas para a disseminação da desinformação do ambiente online provendo uma avaliação diagnóstica que compreenda uma investigação, inclusive, de questões éticas relacionadas.

1.9. Considerações Finais

O presente capítulo apresentou discussões relacionadas ao uso das plataformas digitais para a disseminação de campanhas de desinformação em diferentes contextos. Esperamos que os leitores obtenham uma compreensão aprofundada da desinformação em plataformas digitais e de como combatê-la de forma eficaz, contribuindo assim para um ambiente digital mais saudável e confiável.

Material. Todos os material (i.e., códigos de exemplo, etc) deste capítulo estão disponíveis em um repositório criado para este fim `https://github.com/juliosoaresreis/misinformation-course-jai2023`.

Agradecimentos. Este trabalho foi parcialmente financiado por CNPQ, CAPES, MPMG, FAPEMIG e FAPESP.

[55] `https://openai.com/blog/chatgpt`
[56] `https://openai.com/product/dall-e-2`
[57] `https://midjourney.com/home/`
[58] `https://github.com/CompVis/stable-diffusion`
[59] `https://oglobo.globo.com/economia/tecnologia/noticia/2023/03/midjourney-conheca-a-ferramenta-de-inteligencia-artificial-por-tras-da-foto-do-papa-com-casacao.ghtml`

Referências

Ahmed, H., Traore, I., and Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. In *Proc. of the Int'l Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments (ISDDC)*, pages 127–138.

Allport, G. W. and Postman, L. (1947). The psychology of rumor. *Henry Holt. American Psychological Association*.

Almeida, S. L., Carvalho, P. R., Evangelo, N., and Filgueiras, R. F. D. (2019). Whatsapp: a desordem da informação na eleição presidencial brasileira de 2018. In *Proc. of the Int'l LAVITS*.

Andreou, A., Venkatadri, G., Goga, O., Gummadi, K., Loiseau, P., and Mislove, A. (2018). Investigating ad transparency mechanisms in social media: A case study of facebook's explanations. In *Proc. of the Network and Distributed System Security Symposium (NDSS)*, pages 1–15.

Arun, C. (2019). On whatsapp, rumours, and lynchings. *Econ. & Political Weekly*, 54(6):30–35.

Atanasova, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Karadzhov, G., Mihaylova, T., Mohtarami, M., and Glass, J. (2019). Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–27.

Badawy, A., Addawood, A., Lerman, K., and Ferrara, E. (2019). Characterizing the 2016 russian ira influence campaign. *Social Network Analysis and Mining*, 9(1):31.

Bakshy, E., Messing, S., and Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. (2012). The Role of Social Networks in Information Diffusion. In *The World Wide Web Conf. (WWW'12)*, pages 519–528.

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

Bhattacharjee, S. D., Talukder, A., and Balantrapu, B. V. (2017). Active learning based news veracity detection with feature weighting and deep-shallow fusion. In *Proc. of the IEEE Int'l Conference on Big Data (Big Data)*, pages 556–565.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Budak, C., Goel, S., and Rao, J. M. (2016). Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271.

Castillo, C., Mendoza, M., and Poblete, B. (2011). Information credibility on twitter. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW)*, pages 675–684.

Chakraborty, A., Ghosh, S., Ganguly, N., and Gummadi, K. P. (2016). Dissemination biases of social media channels: On the topical coverage of socially shared news. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 559–562.

Chen, M.-Y., Lai, Y.-W., and Lian, J.-W. (2023). Using deep learning models to detect fake news about covid-19. *ACM Transactions on Internet Technology*, 23(2):1–23.

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., and Flammini, A. (2015). Computational fact checking from knowledge networks. *PLOS ONE*, 10(6):e0128193.

Conroy, N. J., Rubin, V. L., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. In *Proc. of the Annual Meeting of the Association for Information Science and Technology (ASIS&T)*, pages 1–4.

Constantinides, M. (2015). Apps with habits: Adaptive interfaces for news apps. In *Proc. of the Annual ACM Conf. Ext. Abstr. on Human Factors in Comput. Syst. (CHI EA)*, pages 191–194.

Constantinides, M., Dowell, J., Johnson, D., and Malacria, S. (2015). Habito news: A research tool to investigate mobile news reading. In *Proc. of the Int'l Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*, pages 598–598.

Couto, J. M., Pimenta, B., de Araújo, I. M., Assis, S., Reis, J. C., da Silva, A. P. C., Almeida, J. M., and Benevenuto, F. (2021). Central de fatos: Um repositório de checagens de fatos. In *Anais do III Dataset Showcase Workshop (DSW/SBBD)*, pages 128–137.

Couto, J. M., Reis, J. C., Cunha, Í., Araújo, L., and Benevenuto, F. (2022a). Caracterizando websites de baixa credibilidade no brasil. In *Anais do XL Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*, pages 503–516.

Couto, J. M., Reis, J. C., Cunha, Í., Araújo, L., and Benevenuto, F. (2022b). Characterizing low credibility websites in brazil through computer networking attributes. In *Proc. of the IEEE/ACM Intl Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 42–46.

Covert, T. J. A. and Wasburn, P. C. (2007). Measuring media bias: A content analysis of time and newsweek coverage of domestic social issues, 1975–2000. *Social Sci. Quart.*, 88(3):690–706.

Cui, L. and Lee, D. (2020). Coaid: Covid-19 healthcare misinformation dataset.

Cui, L., Shu, K., Wang, S., Lee, D., and Liu, H. (2019). defend: A system for explainable fake news detection. In *Proc. of the Int'l ACM Conference on Information and Knowledge Management (CIKM)*, pages 2961–2964.

da Silva, G. G. (2021). Memes war:: The political use of pictures in brazil 2019. *Philósophos - Revista de Filosofia*, 25(2).

Dai, E., Sun, Y., and Wang, S. (2020). Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 853–862.

Das, A. S., Datar, M., Garg, A., and Rajaram, S. (2007). Google news personalization: scalable online collaborative filtering. In *Proc. of the Int'l ACM Conference on World Wide Web Conference (WWW)*, pages 271–280.

De Angeli, P. and Reis, J. C. (2022). Analyzing the potential of feature groups for misinformation detection in whatsapp. In *Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)*, pages 45–48.

Du, L., Ho, A. T., and Cong, R. (2020). Perceptual hashing for image authentication: A survey. *Signal Processing: Image Communication*, 81:115713.

Ferrara, E. (2017). Disinformation and social bot operations in the run up to the 2017 french presidential election. *First Monday*, 22(8).

Ferrara, E. (2020). What types of covid-19 conspiracies are populated by twitter bots? *First Monday, 25(6)*.

Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7):96–104.

Finn, S., Metaxas, P. T., Mustafaraj, E., O'Keefe, M., Tang, L., Tang, S., and Zeng, L. (2014). Trails: A system for monitoring the propagation of rumors on twitter. In *Computation and Journalism Symposium (C+J)*.

Friggeri, A., Adamic, L. A., Eckles, D., and Cheng, J. (2014). Rumor cascades. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 101–110.

Gallagher, K. (2017). The social media demographics report: Differences in age, gender, and income at the top platforms. http://www.businessinsider.com/the-social-media-demographics-report-2017-8.

Gao, W., Li, P., and Darwish, K. (2012). Joint topic modeling for event summarization across news and social media streams. In *Proc. of the Int'l ACM Conference on Information and Knowledge Management (CIKM)*, pages 1173–1182.

Garcin, F., Galle, F., and Faltings, B. (2014). Focal: A personalized mobile news reader. In *Proc. of the Int'l ACM Conference on Recommender Systems (RecSys)*, pages 369–370.

Garimella, K., Morales, G. D. F., Gionis, A., and Mathioudakis, M. (2018). Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proc. of the Int'l ACM Conference on World Wide Web Conference (WWW)*, pages 913–922.

Garimella, K. and Tyson, G. (2018). Whatapp doc? a first look at whatsapp public group data. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 511–517.

Gentzkow, M., Shapiro, J. M., and Stone, D. F. (2015). Media bias in the marketplace: Theory. 1:623–645.

Golbeck, J., Mauriello, M., Auxier, B., Bhanushali, K. H., Bonk, C., Bouzaghrane, M. A., Buntain, C., Chanduka, R., Cheakalos, P., Everett, J. B., et al. (2018). Fake news vs satire: A dataset and analysis. In *Proc. of the Int'l ACM Conference on Web Science (WebScience)*, pages 17–21.

Gruppi, M., Horne, B. D., and Adalı, S. (2020). Nela-gt-2019: A large multi-labelled news dataset for the study of misinformation in news articles.

Gupta, A., Kumaraguru, P., Castillo, C., and Meier, P. (2014). Tweetcred: Real-time credibility assessment of content on twitter. In *Proc. of the Int'l Conf. on Social Informatics (SocInfo)*, pages 228–243.

Hui, P.-M., Shao, C., Flammini, A., Menczer, F., and Ciampaglia, G. L. (2018). The hoaxy misinformation and fact-checking diffusion network. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 528–530.

Jin, Z., Cao, J., Jiang, Y.-G., and Zhang, Y. (2014). News credibility evaluation on microblog with a hierarchical propagation model. In *Proc. of the IEEE Int'l Conference on Data Mining (ICDM)*, pages 230–239.

Jin, Z., Cao, J., Zhang, Y., Zhou, J., and Tian, Q. (2017). Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia*, 19(3):598–608.

Júnior, M., Melo, P., Kansaon, D., Mafra, V., Sa, K., and Benevenuto, F. (2022). Telegram Monitor: Monitoring Brazilian Political Groups and Channels on Telegram. In *Proc. of the ACM Conference on Hypertext and Social Media (HT)*, page 228–231.

Kansaon, D., Melo, P., and Benevenuto, F. (2022). "click here to join": A large-scale analysis of topics discussed by brazilian public groups on whatsapp. In *Proc. of the Brazilian Symposium on Multimedia and the Web (WebMedia)*.

Keeling, M. J. and Eames, K. T. (2005). Networks and Epidemic Models. *Journal of The Royal Society Interface*, 2(4):295–307.

Kim, J. H., Mantrach, A., Jaimes, A., and Oh, A. (2016). How to compete online for news audience: Modeling words that attract clicks. In *Proc. of the Int'l ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1645–1654.

Kourogi, S., Fujishiro, H., Kimura, A., and Nishikawa, H. (2015). Identifying attractive news headlines for social media. In *Proc. of the Int'l ACM Conference on Information and Knowledge Management (CIKM)*, pages 1859–1862.

Kumar, S., Asthana, R., Upadhyay, S., Upreti, N., and Akbar, M. (2020). Fake news detection using deep learning models: A novel approach. *Transactions on Emerging Telecommunications Technologies*, 31(2):e3767.

Kumar, S., West, R., and Leskovec, J. (2016). Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW)*, pages 591–602.

Kwak, H. and An, J. (2014). A first look at global news coverage of disasters by using the gdelt dataset. In *Proc. of the Int'l Conference on Social Informatics (SocInfo)*, pages 300–308.

Kwon, S., Cha, M., and Jung, K. (2017). Rumor detection over varying time windows. *PloS One*, 12(1):e0168344.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., and Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Lee, C. S. and Ma, L. (2012). News sharing in social media: The effect of gratifications and prior experience. *Computers in human behavior*, 28(2):331–339.

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proc. of the 11th SIGKDD Int'l Conf. on Knowledge Discovery in Data Mining*, pages 177–187.

Li, G. and Zhen, J. (2005). Global stability of an SEI epidemic model with general contact rate. *Chaos, Solitons Fractals*, 23(3):997 – 1004.

Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., and Han, J. (2016). A survey on truth discovery. *ACM SIGKDD Explorations Newsletter*, 17(2):1–16.

Li, Y., Li, Q., Gao, J., Su, L., Zhao, B., Fan, W., and Han, J. (2015). On the discovery of evolving truth. In *Proc. of the Int'l ACM Conf. on Knowl. Disc. and Data Mining (KDD)*, pages 675–684.

Lu, Y.-J. and Li, C.-T. (2020). Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. pages 505–514.

Marques, I., Salles, I., Couto, J. M., Pimenta, B. C., Assis, S., Reis, J. C., da Silva, A. P. C., de Almeida, J. M., and Benevenuto, F. (2022). A comprehensive dataset of brazilian fact-checking stories. *Journal of Information and Data Management*, 13(1).

McAuley, J. and Leskovec, J. (2012). Image Labeling on a Network: Using Social-Network Metadata for Image Classification. In *12th European Conf. on Computer Vision (ECCV12)*.

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)*, pages 79–82.

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

Melo, P., Vieira, C. C., Garimella, K., de Melo, P. O., and Benevenuto, F. (2019b). Can whatsapp counter misinformation by limiting message forwarding? In *Proc. of the Int'l Conference on Complex Networks and their Applications (Complex Networks)*, pages 372–384.

Mitchell, A. (2016). Key findings on the traits and habits of the modern news consumer. `http://www.pewresearch.org/fact-tank/2016/07/07/modern-news-consumer/`.

Monteiro, R. A., Santos, R. L., Pardo, T. A., de Almeida, T. A., Ruiz, E. E., and Vale, O. A. (2018). Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *Proc. of the Int'l Conf. on Comput. Proces. of the Port. Lang. (PROPOR)*.

Müzell, R. B. (2020). Desinformação e Propagabilidade: uma Análise da Desordem Informacional em Grupos de Whatsapp. Master's thesis, Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Rio Grande do Sul.

Nallapati, R., Feng, A., Peng, F., and Allan, J. (2004). Event threading within news topics. In *Proc. of the Int'l ACM Conf. on Inform. and Knowledge Management (CIKM)*, pages 446–453.

Newman, N., Fletcher, R., Kalogeropoulos, A., and Nielsen, R. K. (2019). Reuters Institute Digital News Report 2019. Reuters Institute for the Study of Journalism.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.

Nørregaard, J., Horne, B. D., and Adalı, S. (2019). Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 630–638.

Olson, R. S. and Neal, Z. P. (2015). Navigating the Massive World of Reddit: Using Backbone Networks to Map User Interests in Social Media. *PeerJ Computer Science*, 1:e4.

Oxford (2020). Oxford dictionaries: "memes". https://en.oxforddictionaries.com/definition/meme.

Paul, S., Joy, J. I., Sarker, S., Ahmed, S., Das, A. K., et al. (2019). Fake news detection in social media using blockchain. In *Proc. of the Int'l IEEE Conference on Smart Computing & Communications (ICSCC)*, pages 1–5.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2017). Automatic detection of fake news. pages 3391–3401.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., and Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 231–240.

Poynter (2020). Fighting the infodemic: The coronavirusfacts alliance. https://www.poynter.org/coronavirusfactsalliance/.

Pratiwi, I. Y. R., Asmara, R. A., and Rahutomo, F. (2017). Study of hoax news detection using naïve bayes classifier in indonesian language. In *Proc. of the IEEE Int'l Conference on Information & Communication Technology and System (ICTS)*, pages 73–78.

Quezada, M., Peña-Araya, V., and Poblete, B. (2015). Location-aware model for news events in social media. In *Proc. of the Int'l ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 935–938.

Ratkiewicz, J., Conover, M., Meiss, M. R., Gonçalves, B., Flammini, A., and Menczer, F. (2011). Detecting and tracking political abuse in social media. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 297–304.

Recuero, R., Soares, F., and Vinhas, O. (2021). Discursive strategies for disinformation on whatsapp and twitter during the 2018 brazilian presidential election. *First Monday*.

Reis, J., Benevenuto, F., de Melo, P. O., Prates, R., Kwak, H., and An, J. (2015). Breaking the news: First impressions matter on online news. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 357–366.

Reis, J., Melo, P. d. F., Garimella, K., and Benevenuto, F. (2020a). Can whatsapp benefit from debunked fact-checked stories to reduce misinformation? *The Harvard Kennedy School (HKS) Misinformation Review, 1(5)*.

Reis, J. C. and Benevenuto, F. (2021). Supervised learning for misinformation detection in whatsapp. In *Proc. of the Brazilian Symposium on Multimedia and the Web (WebMedia)*, pages 245–252.

Reis, J. C., Correia, A., Murai, F., Veloso, A., and Benevenuto, F. (2019a). Explainable machine learning for fake news detection. In *Proc. of the ACM Conference on Web Science*, pages 17–26.

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

Reis, J. C., Melo, P., Garimella, K., Almeida, J. M., Eckles, D., and Benevenuto, F. (2020b). A dataset of fact-checked images shared on whatsapp during the brazilian and indian elections. In *Proc. of the Int'l AAAI Conference on Web and Social Media (ICWSM)*, pages 903–908.

Report, D. N. (2018). Statistic of the week: How brazilian voters get their news. https://reutersinstitute.politics.ox.ac.uk/risj-review/statistic-week-how-brazilian-voters-get-their-news.

Resende, G., Melo, P., Sousa, H., Messias, J., Vasconcelos, M., Almeida, J., and Benevenuto, F. (2019). (mis)information dissemination in whatsapp: Gathering, analyzing and countermeasures. In *Proc. of the ACM Web Conference (WWW)*, pages 818–828.

Resende, G., Messias, J., Silva, M., Almeida, J., Vasconcelos, M., and Benevenuto, F. (2018). A system for monitoring public political groups in whatsapp. In *Proc. of the Brazilian Symposium on Multimedia and the Web (WebMedia)*, page 387–390.

Ribeiro, F., Henrique, L., Benevenuto, F., Chakraborty, A., Kulshrestha, J., Babaei, M., and Gummadi, K. P. (2018). Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 290–299.

Ribeiro, F. N., Saha, K., Babaei, M., Henrique, L., Messias, J., Benevenuto, F., Goga, O., Gummadi, K. P., and Redmiles, E. M. (2019). On microtargeting socially divisive ads: A case study of russia-linked ad campaigns on facebook. In *Proc. of the Conference on Fairness, Accountability, and Transparency (FAT)*, pages 140–149.

Ribeiro, M. H., Calais, P. H., Almeida, V. A., and Meira Jr, W. (2017). "everything i disagree with is# fakenews": Correlating political polarization and spread of misinformation. In *Proc. of the Workshop on Data Science + Journalism @KDD*.

Rubin, V., Conroy, N., Chen, Y., and Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news. In *Proc. of the Workshop on Computational Approaches to Deception Detection (NAACL-HLT)*, pages 7–17.

Rubin, V. L., Chen, Y., and Conroy, N. J. (2015). Deception detection for news: three types of fakes. In *Proc. of the Annual Meeting of the Association for Information Science and Technology (ASIS&T)*, page 83. American Society for Information Science.

Ruchansky, N., Seo, S., and Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. In *Proc. of the Int'l ACM Conference on Information and Knowledge Management (CIKM)*, pages 797–806.

Saez-Trumper, D. (2014). Fake tweet buster: a webtool to identify users promoting fake news on twitter. In *Proc. of the ACM Conference on Hypertext and Social Media (HT)*, pages 316–317.

Salem, F. K. A., Al Feel, R., Elbassuoni, S., Jaber, M., and Farah, M. (2019). Fa-kes: A fake news dataset around the syrian war. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 573–582.

Santia, G. and Williams, J. (2018). Buzzface: A news veracity dataset with facebook user commentary and egos. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 531–540.

Sasahara, K., Chen, W., Peng, H., Ciampaglia, G. L., Flammini, A., and Menczer, F. (2020). Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, pages 1–22.

Shao, C., Ciampaglia, G. L., Flammini, A., and Menczer, F. (2016). Hoaxy: A platform for tracking online misinformation. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW) Companion*, pages 745–750.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., and Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications*, 9(1):4787.

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.

Sheidin, J., Lanir, J., Kuflik, T., and Bak, P. (2017). Visualizing spatial-temporal evaluation of news stories. In *Proc. of the Int'l Conference on Intelligence User Interfaces (IUI) Companion*, pages 65–68.

Shu, K., Cui, L., Wang, S., Lee, D., and Liu, H. (2019). defend: Explainable fake news detection. In *Proc. of the Int'l ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 395–405.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Silva, M., Santos de Oliveira, L., Andreou, A., Vaz de Melo, P. O., Goga, O., and Benevenuto, F. (2020). Facebook ads monitor: An independent auditing system for political ads on facebook. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW)*, pages 224–234.

Soares, F. B., Recuero, R., Volcan, T., Fagundes, G., and Sodré, G. (2021). Desinformação sobre o covid-19 no whatsapp: a pandemia enquadrada como debate político. *Ciência da Informação em Revista*, 8(1):74–94.

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., and de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. In *Proc. of the Workshop on Data Science for Social Good (SoGood)*.

Tardaguila, C., Benevenuto, F., and Ortellado, P. (2018). Fake news is poisoning brazilian politics. whatsapp can stop it. `https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html`.

Tomás, R., Tomás, L., and Andreatta, E. (2020). Da depravação ao desperdício de recursos: Estratégias de desconstrução da universidade pública em redes de fake news. *VERBUM. Cadernos de Pós-Graduação.*, 9(2):141–167.

Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., and Krause, A. (2018). Fake news detection in social networks via crowd signals. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW) Companion*, pages 517–524.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323.

Venkatadri, G., Andreou, A., Liu, Y., Mislove, A., Gummadi, K. P., Loiseau, P., and Goga, O. (2018). Privacy risks with facebook's pii-based targeting: Auditing a data broker's advertising interface. In *Proc. of the IEEE Symposium on Security and Privacy (SP)*, pages 89–107.

Vlachos, A. and Riedel, S. (2014). Fact checking: Task definition and dataset construction. In *Proc. of the ACL Workshop on Lang. Technol. and Computat. Social Science*, pages 18–22.

Volkova, S., Shaffer, K., Jang, J. Y., and Hodas, N. (2017). Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 647–653.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

Wang, P., Angarita, R., and Renna, I. (2018a). Is this the era of misinformation yet: combining social bots and fake news to deceive the masses. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW) Companion*, pages 1557–1561.

Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proc. of the Annual Meeting of the Assoc. for Computat. Linguistics (ACL)*, pages 422–426.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., and Gao, J. (2018b). Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proc. of the Int'l ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 849–857.

Wang, Y., Yang, W., Ma, F., Xu, J., Zhong, B., Deng, Q., and Gao, J. (2020). Weak supervision for fake news detection via reinforcement learning. In *Proc. of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 516–523.

Ward, A., Ross, L., Reed, E., Turiel, E., and Brown, T. (1997). Naive realism in everyday life: Implications for social conflict and misunderstanding. *Values and knowledge*, pages 103–135.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'lavits-world' networks. *Nature*, 393(6684):440.

Wei, W. and Wan, X. (2017). Learning to identify ambiguous and misleading news headlines. In *Proc. of the Int'l Joint Conference on Artificial Intelligence (IJCAI)*, pages 4172–4178.

Westlund, O. (2013). Mobile news: A review and model of journalism in an age of mobile media. *Digital journalism*, 1(1):6–26.

Yang, F., Pentyala, S. K., Mohseni, S., Du, M., Yuan, H., Linder, R., Ragan, E. D., Ji, S., and Hu, X. (2019). Xfake: explainable fake news detector with visualizations. In *Proc. of the ACM Web Conference (WWW)*, pages 3600–3604.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of personality and social psychology*, 9(2p2):1.

Zhang, J., Cui, L., Fu, Y., and Gouza, F. B. (2018). Fake news detection with deep diffusive network model. In *arXiv preprint arXiv:1805.08751*.

Zhao, Z., Resnick, P., and Mei, Q. (2015). Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW) Companion*, pages 1395–1405.