Monitor de WhatsApp 2.0
Monitoramento de Grupos Políticos no WhatsApp

Márcio Silva[2], Daniel Kansaon[1], Philipe Melo[3], Fabrício Benevenuto[1]

[1]Universidade Federal de Minas Gerais (UFMG)

[2]Universidade Federal de Mato Grosso do Sul (UFMS)

[3]Universidade Federal de Viçosa (UFV)

Abstract. *WhatsApp has become a crucial tool in communicating and disseminating (mis)information in Brazil. Since 2018, the tool has been widely used for disinformation and hate speech campaigns. In this work, we propose WhatsApp Monitor 2.0, a web-based system that aids researchers and journalists in tracking, in real-time, the most popular content shared in public WhatsApp political groups. Our tool monitors, processes, and ranks images, videos, audios, and text messages posted in these groups, presenting the most popular content daily. WhatsApp Monitor 2.0 provides a valuable resource for identifying viral content on WhatsApp, thus helping to combat misinformation.*

Resumo. *O WhatsApp se tornou uma ferramenta crucial na comunicação e propagação de desinformação no país. Desde de 2018, a ferramenta vem sido amplamente utilizada para campanhas de desinformação e discurso de ódio. Este trabalho, propõe o Monitor de WhatsApp 2.0, um sistema web que auxilia pesquisadores e jornalistas a acompanharem, em tempo real, os conteúdos mais populares compartilhados em grupos públicos do WhatsApp. A ferramenta monitora, processa e ranqueia imagens, vídeos, áudios e mensagens de texto postadas nesses grupos, apresentando diariamente o conteúdo mais popular. O Monitor de WhatsApp 2.0 proporciona um recurso valioso para identificar conteúdos virais no WhatsApp, ajudando no combate à desinformação.*

1. Introdução

Uma parcela substancial da população mundial está ativamente engajada em uma variedade de plataformas digitais e aplicativos de mensagens instantâneas, como o WhatsApp e Telegram [Gallagher 2017]. Estas plataformas têm transformado significativamente a maneira como as pessoas interagem, se comunicam e se mantêm informadas. Nos últimos anos houve um aumento expressivo na quantidade de indivíduos que utilizam redes sociais como fonte de notícias [Newman et al. 2019]. Segundo o IBGE [ASCOM 2021], no Brasil, 82% dos domicílios têm acesso à Internet, e o WhatsApp é o principal meio de informação dos brasileiros [DataSenado 2019], onde 79% dos usuários utilizam o WhatsApp diariamente para compartilhar e discutir notícias com amigos e familiares em conversas dentro da plataforma.

Apesar dos inúmeros benefícios que esses aplicativos de mensagens instantâneas trouxeram para a sociedade, eles também têm sido frequentemente associados às campanhas de desinformação, especialmente em contextos políticos [Resende et al. 2019]. Com

mais de 2 bilhões de usuários, o WhatsApp é atualmente uma das principais plataformas que sofre com desinformação, apontado em vários eventos como pivô do espalhamento de *fake news*, incluindo as campanhas de desinformação durante as eleições presidenciais brasileiras [Kansaon et al. 2024], sobre as urnas eletrônicas [Braun 2022], mudanças climáticas [Santini and Barros 2022], pandemia de COVID-19 [Vijaykumar et al. 2021].

Um dos maiores desafios associados ao WhatsApp é a dificuldade de analisar o conteúdo compartilhado dentro da plataforma. Embora grande parte das conversas ocorra em grupos públicos [Resende et al. 2019], nos quais o conteúdo pode se espalhar rapidamente e alcançar um grande número de pessoas [Melo et al. 2019b] através das funcionalidades que permitem viralização [Melo et al. 2024], a estrutura fechada e a criptografia ponta-a-ponta dificultam a análise deste conteúdo viral. A dificuldade de acesso por meio de pesquisadores e jornalistas torna o desafio de combater à desinformação ainda maior.

Nesse contexto, este trabalho propõe uma nova solução com o *Monitor de WhatsApp 2.0* [1] [2], um sistema web que auxilia os usuários na exploração de conteúdo popular em grupos públicos do WhatsApp. Essa ferramenta expande a versão inicial executada em 2018 [Melo et al. 2019a], incorporando maior acesso e navegação sobre os conteúdos virais de WhatsApp para jornalistas e pesquisadores. Entre as eleições brasileiras de 2018 e dezembro de 2022, jornalistas, pesquisadores e agências de checagem utilizaram a primeira versão do monitor. Diversas notícias se basearam nos dados coletados durante as eleições e a pandemia de COVID-19. Reportagens da BBC [Gragnari 2018, Mori 2019], The Guardian [Avelar 2019], El Pais [Jucá 2019], The Intercept [Pavarin 2019], O Globo [Capetti 2019, Tardáguila 2019b], Folha [Tardáguila 2019a], Uol [UOL 2019], e outras usaram o sistema como referência. Além disso, a UFMG foi parceira do TSE e do MPMG para conter a desinformação nas eleições [TSE 2018].

Já a nova versão do sistema inclui novas implementações que aumentam sua capacidade de fornecer informações sobre a dinâmica dos grupos. O principal diferencial é a utilização de uma arquitetura baseada em microsserviços, que permite o monitoramento em tempo real das mensagens compartilhadas nos grupos. Um painel de tendências exibe as principais mensagens, permitindo também a seleção de um período específico. A interface foi desenvolvida para ser mais intuitiva para o usuário, incluindo um novo sistema de busca otimizado, onde é possível enviar um arquivo e buscar mensagens que contenham esse arquivo. Os conteúdos são processados calculando os *embeddings* dos textos, transcrevendo áudios para texto e avaliando a toxicidade. Essas implementações trazem melhorias significativas, permitindo que pesquisadores e jornalistas acompanhem a dinâmica dos grupos em tempo real, sendo um recurso importante na exploração e monitoramento de ataques e campanhas de desinformação.

2. Arquitetura do Sistema

O conceito inicial do *Monitor de WhatsApp* surgiu em 2018, apresentado em uma versão preliminar no WebMedia'18 [Resende et al. 2018], e posteriormente desenvolvido em um protótipo funcional exibido na ICWSM'19 [Melo et al. 2019a]. O sistema opera com dados provenientes exclusivamente de grupos do WhatsApp que sejam publicamente disponíveis na Web. Para o sistema foram selecionados apenas grupos que abor-

[1]Sistema Disponível em: `https://monitor.semfake.org/`

[2]Coletor disponível em: `https://github.com/LOCUS-DCC-UFMG/monitor_whatsapp`

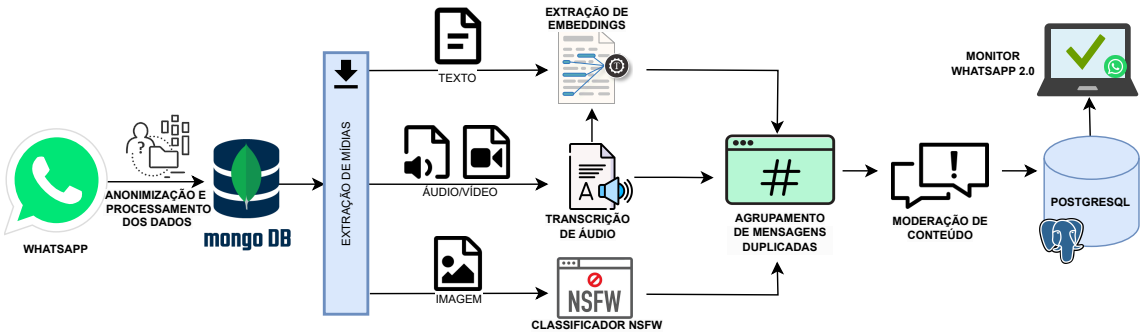

Figure 1. Arquitetura de Coleta e Processamento das Mensagens do WhatsApp.

dam temas políticos, abrangendo um amplo espectro de opiniões, desde discussões mais gerais até grupos mais partidários e/ou extremistas. O acesso a esses grupos é livre por meio de um link de convite específico do WhatsApp: uma URL no formato `chat.whatsapp.com/<groupID>` que pode ser compartilhada em redes sociais para qualquer pessoa que deseje ingressar no grupo.

A nova versão 2.0 passou por uma revisão completa da arquitetura do sistema e da forma como os dados são armazenados. Embora a fonte dos dados permaneça a mesma, houve modificações significativas no acesso e armazenamento dos dados, bem como na arquitetura do sistema em termos de funcionalidades e acesso às informações.

2.1. Coleta e Armazenamento dos Dados

O primeiro passo envolve a seleção dos grupos de interesse a serem monitorados. Para identificar os grupos públicos relevantes, utilizamos uma lista de palavras-chave [Melo et al. 2019a]. Essas palavras-chave são empregadas na busca de grupos públicos em redes sociais, como Twitter e Facebook, bem como na pesquisa em repositórios de grupos públicos por meio do Google. Selecionando apenas os grupos políticos baseados nas palavras-chave, ingressamos nos grupos públicos. Atualmente, são monitorados mais de 1,200 grupos públicos sobre política.

A arquitetura baseada em microsserviços é explicada no fluxograma da Figura 1. Esta figura fornece uma visão geral das etapas executadas pelo sistema online desenvolvido. O gerenciamento dos celulares é feito pela API Evolution[3], um serviço Restful que controla as funções do WhatsApp. Conectando os celulares por meio dessa API, é possível interagir com o WhatsApp utilizando comandos que permitem ler e armazenar as mensagens. Para isso, foi implementado uma API em BUN[4] que interage com a API Evolution. Cada vez que uma nova mensagem é recebida, a API notifica o backend, que inicia uma série de etapas. A cada nova mensagem, a API salva o conteúdo bruto em um banco de dados MongoDB e o ID da mensagem em um banco Redis. O Redis alimenta as filas responsáveis por extrair informações adicionais das mensagens, otimizando o tempo de resposta desses processos. A mensagem é inserida em filas monitoradas por microsserviços, que as leem a cada atualização. Nas filas, são extraídas as mídias, que avaliam se a imagem é segura para ser disponibilizada no monitor. Além disso, são extraídos os *embeddings* dos textos. Já os vídeos e áudios são transcritos para texto e, em seguida, os *embeddings* também são calculados. Uma vez que as mensagens contêm

[3]API: `https://github.com/EvolutionAPI/evolution-api`
[4]BUN Documentation: `https://bun.sh/`

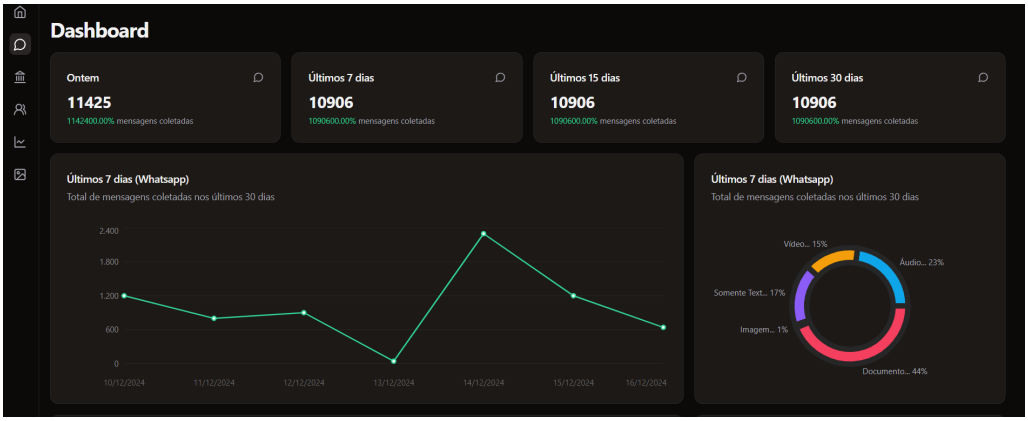

Figure 2. Tela Inicial do Sistema Logo Após o Login.

informações adicionais calculadas, elas são processadas e repassadas para o banco de dados construído em ETL (TimescaleDB - PostgreSQL) focado em colunas temporais, que é acessado pelo monitor. Essa estrutura de coleta e processamento permite a atualização do monitor em tempo real, o que possibilita acompanhar a dinâmica dos grupos, bem como o impacto de eventos em momentos específicos.

2.2. Agrupamento das Mídias

Uma etapa fundamental para o funcionamento do sistema é o agrupamento de conteúdo semelhante. Em redes sociais tradicionais, como Facebook ou Twitter, os posts já vêm com metadados sobre a quantidade de curtidas ou compartilhamentos. Porém, no WhatsApp, cada mensagem é postada de forma independente e isolada. Para contar quantas vezes um determinado conteúdo foi compartilhado, é necessário rastrear e agrupar cada mídia. Para rastrear os compartilhamentos de uma mesma imagem, é utilizado o PerceptualHash (pHash) [Zauner 2010], que calcula um *hash* como uma impressão digital para cada imagem. Para áudios e vídeos, é calculado um *hash* a partir do *checksum* (MD5) de cada arquivo. Com um *hash* para cada mídia, é possível agrupar as postagens do mesmo conteúdo que possuem o mesmo *hash*, permitindo calcular a popularidade, em quantos grupos apareceu e quantos usuários diferentes o enviaram.

2.3. Questões Éticas

Nosso trabalho é restrito apenas a grupos públicos políticos brasileiros. Para assegurar a privacidade dos usuários, não são coletadas informações de identificação pessoal, como números de telefone celular ou nomes de usuários, em conformidade com as normas da LGPD[5]. Os dados no sistema, especialmente as imagens, passam por um filtro de conteúdo[6]) para evitar a exibição de material sensível, violento ou que contenha ódio. Adicionalmente, o sistema é acessível apenas a um número restrito de jornalistas e pesquisadores aprovados, que utilizam contas protegidas por login e senha e preenchem um termo de consentimento sobre a utilização do sistema.

3. Interface e Utilização

O sistema permite aos usuários monitorarem diariamente as tendências compartilhadas em grupos públicos do WhatsApp, exibindo informações sobre mídias de texto, áudio,

[5]Lei Geral de Proteção de Dados. Disponível em: `https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709compilado.htm`

[6]NSFW: `https://github.com/bhky/opennsfw2`

imagem e vídeo compartilhadas grupos. As informações são classificadas do mais popular, com base no alcance observado. Assim que o usuário faz login no sistema, ele recebe a autorização para acessar ao sistema via e-mail. Ao entrar, é mostrado uma visão geral do sistema de coleta e o volume de mensagens nos últimos dias, como mostra a Figura 2.

O usuário pode escolher uma data de início e fim para a pesquisa de conteúdo, onde o sistema recupera e relata o conteúdo mais popular para toda a data selecionada. Isso permite que jornalistas e pesquisadores investiguem um período específico ou até mesmo eventos que durem mais de um dia, combinando milhares de mensagens em uma interface resumida e classificada, na qual alguns padrões de publicação e conteúdo podem emergir. Além de uma interface mais fluida, uma nova funcionalidade é que o sistema permite a busca por meio de mídias. O usuário pode inserir uma imagem, a partir disso, o sistema busca mensagens que contenha o conteúdo, o que permite jornalistas verificarem o quanto um determinado tipo de conteúdo foi compartilhado no WhatsApp. O sistema também oferece uma visão detalhada de cada conteúdo compartilhado. Bastam alguns cliques para revelar o número total de compartilhamentos, quantos grupos participaram da disseminação e quantos usuários únicos compartilharam um conteúdo.

4. Melhorias do Sistema

Os diferenciais do Monitor 2.0 em relação à primeira versão incluem várias melhorias e novas funcionalidades. Primeiramente, a arquitetura baseada em microsserviços e o uso de contêineres Docker proporcionam escalabilidade para o sistema. Além disso, a busca sobre os dados foi ampliada, permitindo filtragem por *query*, data e tipo de mídia (texto, imagem, vídeo, áudio e documento). O sistema possui uma página de tendências, que exibe as principais mensagens em cada tipo de mídia durante um período específico, ordenadas pela quantidade de compartilhamento, seguida pelo número de usuários que compartilharam a mensagem e pela quantidade de grupos. A busca por mídia é uma novidade que permite enviar um arquivo e buscar mensagens que contenham esse arquivo. No caso de imagens, a busca gera dois *hashes* (Phash e PQDHash) que podem ser utilizados para encontrar imagens similares através da distância de cosseno. Para as demais mídias, a busca é feita pelo checksum (sha256). A busca textual utiliza a distância de cosseno para identificar textos semelhantes, calculada a partir dos *embeddings* extraídos via BERT. Além disso, foram implementadas a extração de áudio dos vídeos, a transcrição de áudios para texto com Whisper e a detecção de toxicidade dos textos usando a API de moderação da OpenAI.

Sobre a parte técnica, o sistema implementa controle de permissões baseado em funções Role Based Access Control (RBAC), com três perfis atualmente: Administrador, Público Geral e Órgãos Reguladores. Utiliza um banco de dados ETL (TimescaleDB - PostgreSQL) focado em colunas temporais e com a extensão Pg Vector ativada para busca por *embeddings*. Os dados brutos extraídos diretamente do WhatsApp são armazenados no MongoDB, enquanto o Redis gerencia o processamento das filas.

References

ASCOM (2021). Pesquisa mostra que 82,7% dos domicílios brasileiros têm acesso à internet. *Ministério das Comunicações. gov.br – Governo Federal* . [Online; 14-Abr-2021].

Avelar, D. (2019). Whatsapp fake news during brazil election 'favoured bolsonaro'. `https://shorturl.at/AfSiD`. Accessed on 2024-05-26.

Braun, J. (2022). Conspiração e apuração paralela: a desinformação sobre urnas que circula no whatsapp e telegram às vésperas da eleição. https://www.bbc.com/portuguese/brasil-63097867. Accessed on 2024-05-26.

Capetti, P. (2019). Decisivos na campanha, grupos bolsonaristas no whatsapp agora atuam para desfazer crises. https://shorturl.at/AqZB1. Accessed on 2024-05-26.

DataSenado (2019). Redes Sociais, Notícias Falsas e Privacidade de Dados na Internet. https://shorturl.at/vlssr. Accessed on 2024-09-02.

Gallagher, K. (2017). The social media demographics report differences in age, gender, and income at the top platforms. Insider.

Gragnari, J. (2018). Fighting brazil's election on whatsapp. https://bbc.in/2QFdhMR. Accessed on 2024-05-26.

Jucá, B. (2019). Mobilização por educação confronta bolsonaristas nas redes e testa força nas ruas. https://shorturl.at/ZyRmq. Accessed on 2024-05-26.

Kansaon, D., Melo, P. d. F., Zannettou, S., Feldmann, A., and Benevenuto, F. (2024). Strategies and attacks of digital militias in whatsapp political groups. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):813–825.

Melo, P., Hoseini, M., Zannettou, S., and Benevenuto, F. (2024). Don't break the chain: Measuring message forwarding on whatsapp. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):1054–1067.

Melo, P., Messias, J., Resende, G., Garimella, K., Almeida, J., and Benevenuto, F. (2019a). Whatsapp monitor: A fact-checking system for whatsapp. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):676–677.

Melo, P., Vieira, C. C., Garimella, K., de Melo, P. O. V., and Benevenuto, F. (2019b). Can whatsapp counter misinformation by limiting message forwarding? In *International Conference on Complex Networks and Their Applications*, pages 372–384. Springer.

Mori, L. (2019). Por que convocação de ato pró-bolsonaro está rachando a direita. https://bbc.in/2QFdhMR. Accessed on 2024-05-26.

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

Pavarin, G. (2019). Como a milícia digital bolsonarista resgatou sua máquina de fake news para atacar universitários. https://shorturl.at/vlssr. Accessed on 2024-05-26.

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 *The World Wide Web Conference*, WWW '19, pages 818–828. ACM.

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

Santini, R. M. and Barros, C. E. (2022). Negacionismo climático e desinformação online: uma revisão de escopo. *Liinc em Revista. Desafios das Ciências sociais no Antropoceno*, 18(1):e5948.

Tardáguila, C. (2019a). Fotos (velhas) de universitários nus inundam whatsapp para 'provar' a 'balbúrdia' apontada por weintraub. https://shorturl.at/scV0r. Accessed on 2024-05-26.

Tardáguila, C. (2019b). Marielle, suzano e stf transformaram o whatsapp num pântano de horror e ódio. https://shorturl.at/qiU9b. Accessed on 2024-05-26.

TSE (2018). Tse estuda possibilidade de firmar parceria com universidade para inibir fake news no whatsapp. https://shorturl.at/JgCWu. Accessed on 2024-07-08.

UOL (2019). Para inflar economia do brasil, áudio inventa que mercedes esgotou produção. https://shorturl.at/u2Ilq. Accessed on 2024-05-26.

Vijaykumar, S., Jin, Y., Rogerson, D., Lu, X., Sharma, S., Maughan, A., Fadel, B., de Oliveira Costa, M. S., Pagliari, C., and Morris, D. (2021). How shades of truth and age affect responses to covid-19 (mis) information: randomized survey experiment among whatsapp users in uk and brazil. *Humanities and Social Sciences Communications*, 8(1):88.

Zauner, C. (2010). Implementation and benchmarking of perceptual image hash functions.