

A System for Monitoring Public Political Groups in WhatsApp

Gustavo Resende
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais
gustavo.jota@dcc.ufmg.br

Johnnatan Messias
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais
johnnatan@dcc.ufmg.br

Márcio Silva
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais
marcio@facom.ufms.br

Jussara Almeida
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais
jussara@dcc.ufmg.br

Marisa Vasconcelos
IBM Research
São Paulo, São Paulo
marisaav@br.ibm.com

Fabício Benevenuto
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais
fabricao@dcc.ufmg.br

ABSTRACT

In Brazil, 48% of the population use WhatsApp to share and discuss news. Currently, there are serious concerns that this platform can become a fertile ground for groups interested in disseminating misinformation, especially as part of articulated political campaigns. Particularly, WhatsApp provides an important space for users to engage in public conversations that worth attention, the public groups. These groups are suitable for political activism and social movement organization. Additionally, it is reasonable to assume that a malicious misinformation campaign might attempt to maximize the audience of a fake story by sharing it in existing public groups. In this paper, we present a system for gathering, analyzing and visualize public groups in WhatsApp. In addition to describe our methodology, we also provide a brief characterization of the content shared in 127 Brazilian groups. We hope our system can help journalists and researchers to understand the repercussion of events related to the Brazilian elections within these groups.

CCS CONCEPTS

• **Information systems** → **Data analytics**; *Social networking sites*; *Mobile information processing systems*; *Chat*; *Texting*;

KEYWORDS

WhatsApp, Politics, Chat, App, Mobile, Groups

ACM Reference Format:

Gustavo Resende, Johnnatan Messias, Márcio Silva, Jussara Almeida, Marisa Vasconcelos, and Fabício Benevenuto. 2018. A System for Monitoring Public Political Groups in WhatsApp. In *Brazilian Symposium on Multimedia and the Web (WebMedia '18)*, October 16–19, 2018, Salvador-BA, Brazil. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3243082.3264662>

1 INTRODUÇÃO

Desde seu surgimento, o *WhatsApp* tem mudado a forma de comunicação dos usuários de *smartphones*, por permitir uma interação

mais fluida nas conversas. Com a facilidade de compartilhamento de diversos tipos de mídia como imagens, áudios e vídeos de maneira instantânea, o aplicativo se tornou extremamente popular, chegando a mais de 1.5 bilhão de usuários ativos atualmente [4]. Outro aspecto do aplicativo é a criação de grupos na qual os usuários podem trocar mensagens simultaneamente. Esses grupos podem ser privados, se o ingresso de novos membros é feito por intermédio de um membro administrador do grupo, ou *públicos* acessíveis por meio de links de convite que são publicados em sites conhecidos, bem como em diversas redes sociais como *Facebook* e *Twitter* e normalmente possuem temas de discussão específicos como política, educação.

O *WhatsApp*, tal qual em outras ferramentas sociais como o *Twitter* [1], tem sido apontado como uma das principais ferramentas de disseminação de conteúdo na atualidade, conforme observado na greve dos caminhoneiros em Maio de 2018, em que possibilitou a mobilização de milhares deles[2]. Outro fato recente, foi a disseminação de rumores sobre sequestro de crianças na Índia [9].

Ainda existem poucos estudos sobre a utilização desse aplicativo, muitos deles através de usuários voluntários que disponibilizaram suas mensagens [6] ou através da monitoração das mensagens e das sessões de usuários em grupos públicos políticos e não políticos [3, 5]. Porém, esses estudos não analisaram as mensagens de mídia separadamente, do envio de imagens, áudios e vídeos.

Neste artigo, apresentamos um sistema para coletar, analisar e visualizar grupos públicos no WhatsApp. Além de descrever nossa metodologia, também fornecemos uma breve caracterização do conteúdo compartilhado por 6.314 usuários em 127 grupos públicos brasileiros do *WhatsApp* com temáticas relacionadas à política e notícias gerais. Nós acreditamos que nosso sistema possa ajudar jornalistas e pesquisadores a entender a repercussão de eventos relacionados às eleições brasileiras dentro desse espaço midiático.

2 TRABALHOS RELACIONADOS

Poucos trabalhos exploraram o uso do WhatsApp na literatura. Alguns estudos monitoraram usuários do WhatsApp[6, 8], que voluntariamente permitiram a coleta de dados ou estatísticas como tempo médio de uso do aplicativo. Rosenfeld *et al.* [8] realizaram uma análise do comportamento de 100 usuários para inferência de informações demográficas dos usuários como idade e gênero, já Montag *et al.* [6] monitoraram 2.418 usuários e observaram que o *WhatsApp* era responsável por 20% do tempo de uso do *smartphone*, enquanto que as mulheres utilizam o aplicativo mais tempo que os homens. O uso do aplicativo para relato de incidentes durante a

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WebMedia '18, October 16–19, 2018, Salvador-BA, Brazil

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-5867-5/18/10...\$15.00

<https://doi.org/10.1145/3243082.3264662>

eleição em Gana em 2016, foi analisado em [7]. Enquanto, os autores citados focaram apenas no texto das mensagens, este trabalho, analisa vários grupos públicos do *WhatsApp* focados em política e além disso caracteriza o compartilhamento de outros tipos de conteúdos como imagens, áudios e vídeos.

Alguns trabalhos recentes apresentam metodologias de coletas de dados de grupos públicos no *WhatsApp* e provêm análises complementares às nossas [3, 5]. Diferentemente, nosso trabalho apresenta um sistema e permite o monitoramento da repercussão de eventos nesse espaço do *WhatsApp*.

3 MONITOR DE GRUPOS PÚBLICOS

Nesta seção são apresentados os principais componentes do nosso sistema de visualização de grupos públicos do *WhatsApp*¹. Como entrada do sistema são coletados diversos tipos de mídia disponíveis (texto, áudio, imagem e vídeo) por 127 grupos públicos relacionados a discussão de temas políticos e de notícias. Em seguida é feito um processo de seleção das mídias mais populares e categorização dessas para a posterior visualização. O objetivo desse sistema é informar e antecipar comunicadores sobre o tipo de informação compartilhada pelos brasileiros em grupos públicos. A ideia é dar acesso ao sistema para profissionais provedores da informação e agentes de segurança pública, que são capazes de realizar a checagem de fatos e informações compartilhadas em grupos públicos. Ressaltamos ainda que não é disponibilizado, nenhuma informação pessoal dos membros do grupo e como descrito nas seções abaixo não violamos a política de privacidade do *WhatsApp*.

3.1 Coleta de Dados

Assim como Caetano et al., foi utilizada também uma lista de palavras-chaves² relacionadas à política e a notícias juntamente com o link “*chat.whatsapp.com*” para coleta dos grupos públicos do *Whatsapp* no *Google*, *Twitter* e *Facebook*. Foram obtidos 3.447 links distintos para grupos públicos, dos quais 1.828 eram válidos.

Para a subscrição automática de grupos coletados, foi utilizado o mesmo script de [5] que recebe como entrada a lista de links de acesso a esses grupos. Os grupos cujos nomes se relacionavam com os termos da lista de palavras-chaves mencionada acima foram coletados. Uma vez que a busca também retorna links para diversos grupos que não se relacionam com os termos buscados, dos 1.828 grupos válidos verificados, 127 grupos estavam de acordo com a temática das palavras chaves e foram vinculados e divididos entre 3 aparelhos *smartphones* utilizados para coleta.

Uma vez a subscrição nos grupos concluída nos 3 aparelhos celulares, o processo de coleta é iniciado. Em razão da criptografia de ponta-a-ponta utilizada nas mensagens, foi utilizada também a ferramenta *WebWhatsAppAPI*³ que coleta as mensagens utilizando a versão web do *WhatsApp*.

Durante o período 27/04 a 30/05 de 2018, foram coletadas todas as mensagens compartilhadas em 127 grupos públicos. De cada mensagem foram extraídas informações como nome do grupo, ID do grupo, ID do usuário, data e horário. Para mensagens de mídia,

foram armazenados também os arquivos de áudio, vídeo, imagem e o nome do arquivo para referência a mensagem. No total foram coletadas um volume de 210.609 mensagens, 169.154 mensagens de texto, 5.999 mensagens de áudio, 14.324 vídeos e 21.132 imagens.

Dentre as imagens coletadas, foram filtradas aquelas classificadas como adultas ou ofensivas utilizando-se um modelo de detecção de imagens NSFW (*Not Safe For Work*)⁴. O modelo retorna uma probabilidade (entre 0 e 1) da imagem ser imprópria. Para esse trabalho, imagens com valores acima de 0.8, como sugerido pela ferramenta, foram classificadas como de conteúdo impróprio. Durante o período monitorado, 0.6% das imagens foram classificadas como impróprias.

3.2 Limitações dos Dados e Termos de Serviço

Esse trabalho se baseia apenas na análise de grupos públicos do *WhatsApp*, que são os grupos possíveis de serem coletados. O *WhatsApp* permite que cada grupo possua no máximo 256 membros. Diante desse cenário, não há conhecimento sobre o quanto do todo a base de dados representa. Porém, esse trabalho é o primeiro a analisar o debate político dentro desse sistema através da quantificação do viés ideológico e representatividade por estado brasileiro.

A Política de Privacidade do *WhatsApp*⁵ define que todas as informações do perfil do usuário como nome de usuário e telefone celular estão disponíveis para qualquer outro usuário que tenha uma interação pelo aplicativo, seja uma conversa privada ou em grupo, estando estas informações disponíveis para todos os membros em grupos sejam eles privados ou públicos. Para manter a privacidade dos usuários, os dados relativos ao nome do usuário e ao número de celular foram devidamente anonimizados sendo mantidos somente os códigos de discagem direta à distância, para análise geográfica.

4 CARACTERIZAÇÃO DOS DADOS

Essa seção caracteriza o viés político (Seção 4.1), demográfico (Seção 4.2) e o tipo de conteúdo compartilhado (Seção 4.3).

4.1 Viés Político dos Grupos Públicos

Como descrito na Seção 3, foi feita uma busca por grupos públicos com temáticas relacionadas à política, notícias. De forma a estudar a interação, foram criadas as seguintes categorias baseadas no nome do grupo e no conteúdo das mensagens: Notícias, Debates, Direita, Ideologias, Pró-Lula, Pró-Bolsonaro e Partidos.

Os grupos da categoria *Notícias* se caracterizam por mensagens que compartilham e discutem notícias que podem vir na forma de links, mensagens de texto ou imagens. Os grupos de *Debates* discutem principalmente sobre política e corrupção. Já na categoria *Direita* estão os grupos que discutem o aspecto político chamado de Direita e na categoria *Ideologia* estão grupos que seguem alguma ideologia específica (ex.anarquismo, imperialismo e ateísmo).

Categoria	Grupos	Texto	Imagens	Áudios	Vídeos
Debates	26	68.523	6.324	1.515	3.910
Direita	8	9.165	1.155	299	917
Ideologias	21	36.716	1829	382	1.064
Notícias	30	40.416	7.228	2.460	4617
Partidos	15	6.272	1558	369	1.051
Pro-Bolsonaro	18	4.922	1937	816	2.142
Pro-Lula	9	3.140	1101	158	623

Tabela 1: Distribuição de Mensagens por Categoria

¹<http://www.monitor-de-whatsapp.dcc.ufmg.br>

²<https://docs.google.com/document/d/1T3U6QJQVvGy4Wh-3uLDbCG9IeX0d4KbMpfzTui1TC4U/edit?usp=sharing>

³<http://github.com/mukulhase/WebWhatsApp-Wrapper>

⁴https://github.com/yahoo/open_nsfw

⁵https://www.whatsapp.com/legal/?lang=pt_br#privacy-policy

Foi observado, que um número significativo de grupos públicos é dedicado a políticos específicos, em particular grupos que representam o interesse no ex-presidente Lula e no pré-candidato das eleições de 2018 Jair Bolsonaro. Assim, foram criados duas categorias que abrangem esses dois políticos: *Pro-Lula* e *Pro-Bolsonaro*. Grupos dedicados a outros partidos e candidatos específicos (ex. Ciro Gomes e Partido Novo) foram agrupados na categoria *Partidos*. A Tabela 1 apresenta a distribuição do número de mensagens por categoria dos grupos. Em razão do número maior de grupos públicos coletados, as categorias *Debates* e *Notícias* apresentaram também o maior volume de mensagens de texto e de mídia.

Foram analisadas também as palavras mais comumente usadas pelos usuários nos grupos *Pró-Bolsonaro* e *Pró-Lula*. A Figura 1 apresenta as nuvens de palavras para os dois políticos. Para os grupos *Pró-Bolsonaro* (Figura 1-a), a palavra como maior destaque foi *Bolsonaro*, o que reforça a temática principal desses grupos, além de termos como *governo*, *Brasil*, que fazem sentido uma vez que se trata seguidores desse candidato a presidente do Brasil. Termos como *greve*, *caminhoneiro* e *gasolina* também foram frequentes mostrando a discussão do evento da greve dos caminhoneiros nesses grupos.

Já para os grupos *Pró-Lula* (Figura 1-b), observou-se termos como *Lula*, *Sérgio Moro*, *Lula Livre*, *companheiro* sendo os mais comuns. Apesar de não terem sido encontrados grupos públicos cuja temática principal era referente ao posicionamento político chamado de Esquerda, os grupos da temática *Pró-Lula*, podem ser considerados de Esquerda, uma vez que o ex-presidente Lula que faz parte da temática principal destes grupos apresenta um posicionamento político voltado para a chamada Esquerda.



(a) Pró-Bolsonaro

(b) Pró-Lula

Figura 1: Palavras mais usadas nas mensagens dos grupos

4.2 Viés Demográfico dos Grupos

Os grupos públicos, por serem abertos e não necessitarem de uma aprovação para o ingresso de novos membros, tendem a possuir membros de diversas regiões. Para entender como os usuários dos grupos públicos do *WhatsApp* se distribuem geograficamente por região brasileira, foi feita uma classificação de cada grupo utilizando-se o DDD mantido de cada usuário. Foram classificados como grupos locais aqueles em que a maioria de seus membros (mais de 50%) são residentes do mesmo estado (ou seja, possuem o mesmo DDD). Os grupos locais com um maior número de membros, são compostos por usuários do Rio de Janeiro, seguido por Goiás e São Paulo.

Na Figura 2(a) são listados os 15 Estados com maior número de postagens nos grupos monitorados, onde o Estado de São Paulo lidera o número de mensagens enviadas, seguido pelo Rio de Janeiro e Bahia. Ceará e Goiás possuem uma participação semelhante. É notável que há a presença de Estados de todas as regiões do Brasil.

Além disso verificamos a distribuição em termo de usuários, contidos em nossa base. Foram identificados 6314 usuários únicos. A Figura 2(b) lista os 15 estados com maior número de usuários

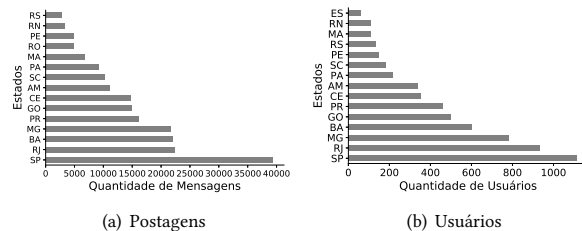


Figura 2: Distribuição por Estado Brasileiro

onde o estado de São Paulo, seguido por Rio de Janeiro e Bahia são os estados com maior número de usuários o que justifica eles também terem o maior número de mensagens como na Figura 2(a).

4.3 Tipos de Mídia Compartilhada

Nessa seção, são analisados outros conteúdos (links, imagens, vídeos e áudios) compartilhados pelos grupos públicos coletados.

URLs e Domínios: Observou-se que 25.491 mensagens compartilhadas pelos grupos públicos possuíam URLs, representando por volta de 10% do total de mensagens de texto coletadas. Nas mensagens de texto foram encontrados 29.110 links onde 20.608 são únicos, uma vez que uma mensagem pode conter mais de uma URL. Foram encontrados 1.564 domínios diferentes nas URLs presentes nas mensagens compartilhadas. A Figura 3 mostra os 10 domínios mais populares presentes nas mensagens coletadas. Pode se observar que os domínios mais populares são do *YouTube* e do *Facebook*, o que mostra que esses usuários compartilham muito conteúdo de outras redes sociais e conteúdo de vídeo. Os outros domínios presentes na base são referentes a portais de notícias locais e nacionais e blogs de conteúdo político.

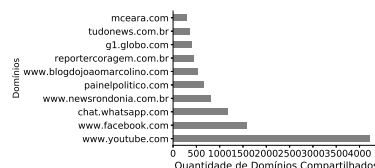


Figura 3: Distribuição dos 10 domínios mais compartilhados

Imagens: Imagens representam o tipo de mídia mais popular nos grupos públicos monitorados do *WhatsApp*, quase 10% do total de mensagens enviadas. Das mensagens obtidas 21.132 são imagens e 1.331 são imagens duplicadas ou repetidas. A Figura 4 apresenta o histograma de mensagens enviadas por dia. Nota-se que em grande maioria dos dias monitorados foram enviadas cerca de 600 imagens. Em pouco dias o número de imagens enviadas foi maior que 1.000, podendo ser um indicador que se tratam de dias atípicos.

Áudios: Nos grupos monitorados, os áudios foram o tipo de mídia menos compartilhado (Tabela 1 e Figura 4). Durante a coleta foram enviados 5.999 áudios, sendo 269 duplicados ou repetidos representando 3% das mensagens, com cerca de 100 áudios por dia. Em alguns dias houveram mais 300 áudios compartilhados, em que verificou-se que se tratava do período da greve dos caminhoneiros.

Vídeos: Nas mensagens dos grupos monitorados foram enviados 14.324 vídeos no total, sendo 642 duplicados ou repetidos representando 6.8% das mensagens monitoradas. No histograma da Figura 4, é notável que o comportamento mais frequente é o envio de cerca de 200 vídeos por dia, verificado em 44% dos dias monitorados. Em

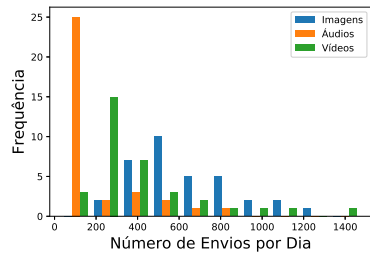


Figura 4: Mídias enviadas por dia

seguida em 29% dos dias foram enviados aproximadamente 400 vídeos. Em poucos dias o número de vídeos enviados se apresentou maior que 1.000, porém assim como nos áudios, tal comportamento ocorreu durante a greve dos caminhoneiros.

5 ESTUDO DE CASO

O *WhatsApp* tem tido uma grande importância na mobilização social de grandes eventos e acontecimentos. Dois eventos monitorados em 2018 foram destacados no sistema pelo volume de mensagens e mídias compartilhadas, uma foi a prisão do ex-presidente Lula (Abril) e o outro foi a greve dos caminhoneiros (Maio).

O sistema descrito aqui exibe quais conteúdos são os mais compartilhados nos grupos públicos monitorados. A Figura 5 destaca as imagens mais compartilhadas em dois dias da coleta. Pode-se observar a presença de imagens relacionadas a prisão do ex-presidente e o início do movimento *Lula Livre*, que se iniciou após a prisão, como pode ser visto na Figura 5(a) em que o termo foi escrito em cédulas de reais. Já as imagens mais populares postadas no dia 21 de maio estão relacionadas a greve nacional dos caminhoneiros, que se mobilizaram utilizando o próprio *WhatsApp*. No período da greve os grupos monitorados tiveram um comportamento diversificado. Como pode ser visto nas Figuras 1(a) e 1(b) as palavras *Caminhoneiros*, *Gasolina* e *Greve* foram identificadas na nuvem de palavras de ambas as categorias de grupo Pró-Lula e Pró-Bolsonaro.



(a) 08/04/2018 (b) 21/05/2018

Figura 5: Imagens exibidas pelo sistema

O sistema monitor dos grupos identifica as imagens, áudios e vídeos mais compartilhados por dia. Durante os dias que ocorreram a greve dos caminhoneiros, o número de compartilhamentos dos arquivos de mídia virais de cada dia sofreu um aumento drástico, o gráfico da Figura 6, mostra o número de vezes que mídia mais compartilhada foi enviada. O número de envios do áudio mais compartilhado por dia dobrou no início da greve, no dia 21, chegando ao pico nos dias finais. Já em relação ao vídeo mais compartilhado por dia, o aumento se iniciou a partir do dia 25, mas também foi significativo. Quanto as imagens virais, durante os dias 22 e 24, o número de compartilhamentos da imagem mais enviada teve um pico que foi destoante do usual, maior e mais longo.

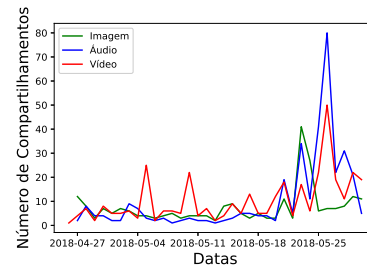


Figura 6: Compartilhamento das mídias por dia

Esses resultados mostram que o conteúdo compartilhado durante a greve dos caminhoneiros teve um impacto perceptível na comunicação dos grupos monitorados pelo sistema. O conteúdo exibido pelo sistema descrito, foi utilizado por jornalistas durante o evento da greve descrita, em matérias publicadas em portais de notícias de renome no Brasil, como a BBC, a Folha de São Paulo e o G1.

6 CONCLUSÃO

Este trabalho apresentou um sistema de monitoração de grupos públicos de *WhatsApp* e uma caracterização do conteúdo compartilhado por 127 grupos públicos de *WhatsApp* relacionados a temas como política e notícias. Do período analisado nesse trabalho, pode-se observar a presença de grande eventos de nível nacional como a greve dos caminhoneiros no Brasil em Maio de 2018 dentre os mais populares. Durante o mesmo evento, o sistema se mostrou útil sendo utilizado por jornalistas e foram publicadas reportagens sobre a greve utilizando dados do sistema em diferentes portais de grande visibilidade nacional, incluindo BBC, G1 e Folha de São Paulo [2]. Tais reportagens reforçam nossa hipótese de que o monitoramento do *WhatsApp* podem trazer informações que são relevantes sobre questões políticas ou de impacto nacional. Atualmente, nós já estamos monitorando mais de 200 grupos públicos e compartilhamos o nosso sistema com mais de 50 jornalistas brasileiros.

ACKNOWLEDGMENTS

Este trabalho foi parcialmente apoiado pelo projeto FAPEMIG-PRONEX-MASWeb, Modelos, Algoritmos e Sistemas para a Web, processo:APQ-01400-14 e por auxílio do CNPq, CAPES e Fapemig.

REFERÊNCIAS

- [1] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. *First Monday* 21, 11 (2016).
- [2] BBC Brasil. 2018. Como o WhatsApp mobilizou caminhoneiros, driblou governo e pode impactar eleições. <https://www.bbc.com/portuguese/brasil-44325458>. Acessado: 09/08/2018.
- [3] Josemar Caetano, Jaqueline de Oliveira, Helder Lima, Humberto Marques-Neto, Gabriel Magno, Wagner Meira Jr, and Virgílio Almeida. 2018. Analyzing and characterizing political discussions in WhatsApp public groups. *arXiv preprint arXiv:1804.00397* (2018).
- [4] Josh Constine. 2018. WhatsApp Hits 1.5 billions monthly users. <https://techcrunch.com/2018/01/31/whatsapp-hits-1-5-billion-monthly-users-19b-not-so-bad/> [Acessado em 31/01/2018].
- [5] Kiran Garimella and Gareth Tyson. 2018. WhatsApp, Doc? A First Look at WhatsApp Public Group Data. In *Proc. of the ICWSM*.
- [6] Christian Montag, Konrad Błazszkiewicz, Rayna Sariyska, Bernd Lachmann, Ionut Andone, Boris Trendafilov, Mark Eibes, and Alexander Markowetz. 2015. Smartphone usage in the 21st century: who is active on WhatsApp? *BMC research notes* 8, 1 (2015), 331.
- [7] Andrés Moreno, Philip Garrison, and Karthik Bhat. 2017. WhatsApp for Monitoring and Response during Critical Events: Aggie in the Ghana 2016 Election. In *ISCRAM*.
- [8] Avi Rosenfeld, Sigal Sina, David Sarne, Or Avidov, and Sarit Kraus. 2018. WhatsApp Usage Patterns and Prediction Models. (2018).
- [9] BBC World. 2018. India 'WhatsApp child abduction rumours'. <https://www.bbc.com/news/world-asia-india-44678674>. Acessado em 09/08/2018.