Análise de engajamento no Facebook em notícias sobre a COVID-19

César Augusto Julio da Silva

Instituto de Computação
Universidade Federal do Rio de Janeiro (UFRJ) – Rio de Janeiro, RJ – Brasil

cesarsilva06@dcc.ufrj.br

Abstract. *The Covid-19 pandemic forced modern society to use virtual media as the main form of communication during 2020 and 2021. During this period, journalistic and informational companies intensified their efforts in this virtual world, trying to promote a much greater engagement by part of the citizens. The focus of this article is to expose engagements throughout the year 2020 within the Facebook platform with a focus on finding out what happened.*

Resumo. *A pandemia de Covid-19 forçou a sociedade moderna a utilizar meios virtuais como a principal forma de comunicação durante o ano de 2020 e 2021. Durante esse período, companhias jornalísticas e informacionais intensificaram seus esforços nesse mundo virtual, tentando promover um engajamento muito maior por parte dos cidadãos. O foco desse artigo é expor engajamentos ao longo do ano de 2020 dentro da plataforma Facebook com o foco de averiguar o que ocorreu.*

1. Introdução

Em 2020, o mundo presenciou uma das maiores crises na saúde da história contemporânea: a pandemia de Covid-19. Essa doença que começou a se proliferar pelo oriente, se alastrou pelo planeta em pouco menos de seis meses afetando interações sociais, economias e políticas públicas e forçando a humanidade a aprender meios alternativos de convivência.

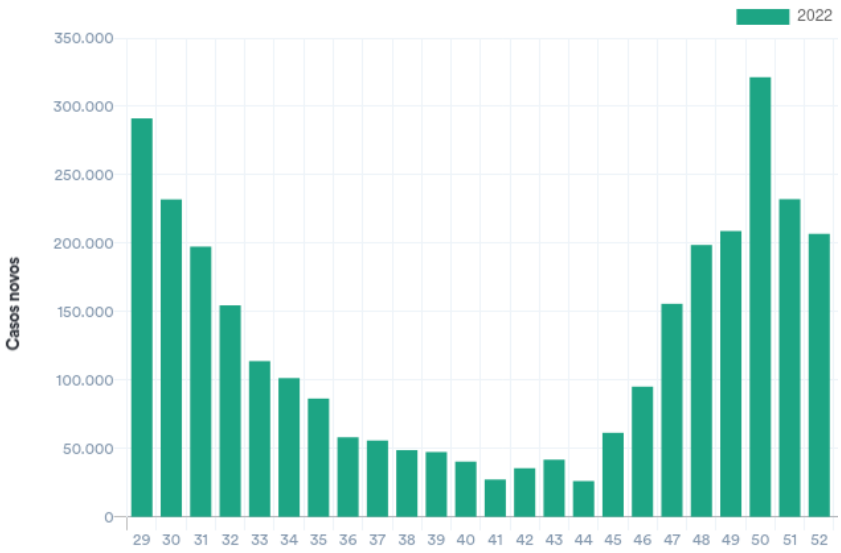

Figura 1. Casos novos de COVID-19 por Semana Epidemiológica de notificação em 2022.

Dentre esses novos desdobramentos, um dos mais intensificados foi a interação virtual por meio das mídias sociais, que apesar de já existir faz alguns anos, obteve uma renovada força durante esse período.

Com isso, muitas empresas aumentaram a quantidade de recursos investidos na divulgação via tais meios midiáticos. Nesse projeto, o foco será no engajamento relacionado a sites de notícias divulgando a própria pandemia.

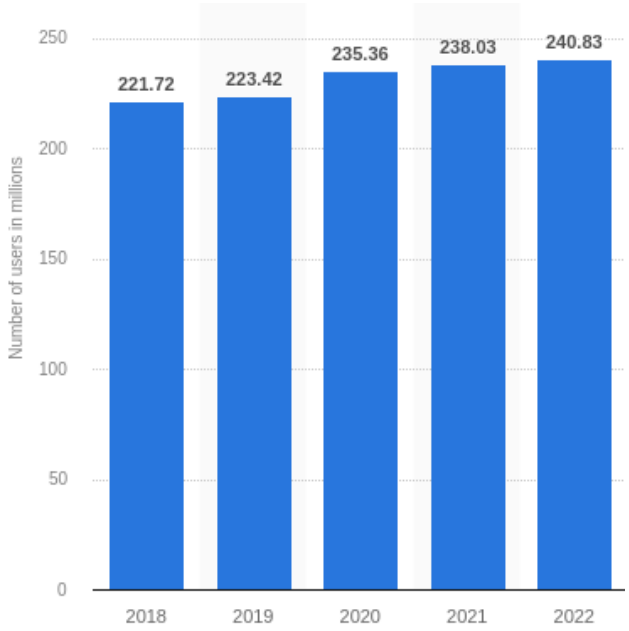

Figura 2. Gráfico da quantidade de usuários do Facebook nos EUA durante os anos de 2018 a 2022.

Inicialmente, o foco escolhido foi responder perguntas como: quais os tipos de notícias atraíram mais atenção, quais sites tiveram mais engajamento dado um mesmo contexto, como foi a reação das pessoas dado um tipo de notícia. Uma adaptação, contudo, foi feita. Para responder tais questionamentos iriam ser escolhidos cinco períodos durante o ano de 2020 e quatro websites de notícias que divulgaram tais momentos, porém muitas das notícias não relataram o mesmo assunto e assim ficaria difícil compará-las entre si. Assim, foi pega uma amostra de 4 notícias a cada mês de 2020, distribuídas entre os quatro domínios e foi analisado o engajamento de tais notícias na plataforma Facebook.

O relatório consiste de uma revisão da literatura onde serão relatados trabalhos semelhantes, materiais utilizados e métodos empregados, resultados das análises e suas avaliações, e uma conclusão final.

2. Revisão de Literatura

Para realizar essa pesquisa, inicialmente averiguamos trabalhos similares feitos anteriormente. Destes, três se aproximaram da proposta desse relatório. Um deles [Massarani, L., et al. 2021], trata sobre o fenômeno de “infodemia”, a circulação de um

grande volume de informações, em parte enganosas ou falsas e os efeitos que ela teve sobre a identificação de fontes confiáveis, em especial, com relação a medidas de contenção, como as vacinas. Em outro estudo [Qiang, C., et al. (2020)], os pesquisadores investigam como as agências centrais do governo chinês usam as mídias sociais para promover engajamento dos cidadãos durante a crise da Covid-19.

2.1. Infodemia, desinformação e vacinas

No primeiro projeto referido, a metodologia do estudo envolveu três etapas: a coleta do corpus, a análise por categorias, e a comparação dos resultados com os dados de 2018 - 2019. A coleta dos links (um total de cem links ao todo) foi realizada usando o BuzzSumo, uma ferramenta de monitoramento e análise de conteúdo. Infelizmente, a ferramenta é paga então não foi viável utilizá-la para esse projeto.

Em seguida, foram definidas cinco categorias de análise: engajamento, tema, acurácia, posicionamento, e tipo de veículo; todas baseadas em pesquisas anteriores [Massarani et al. 2020]. Essas categorias não foram utilizadas devido aos metadados fornecidos pela ferramenta utilizada nesse projeto, CrowdTangle, não serem iguais aos do BuzzSumo, além do enfoque de busca ser somente na rede social Facebook, e em engajamento, não em veracidade das informações. Por último, foram comparados os resultados referentes a 2020 aos obtidos a partir da análise do corpus 2018-2019.

Os resultados apontaram um aumento no engajamento médio em 8,6 vezes e que a predominância de informações verificadas se manteve antes e durante a pandemia. No entanto, também revelou que o engajamento da desinformação cresceu de maneira expressiva e seu perfil mudou: se em 2018-2019 predominavam os conteúdos totalmente falsos e emitidos por veículos não profissionais, já em 2020 se destacam as informações distorcidas por manchetes sensacionalistas emitidas por veículos profissionais. Além disso, também testemunha a instrumentalização política do debate sobre vacinação mostrando uma forte relação entre desinformação e disputas narrativas.

2.2. Promover engajamento dos cidadãos durante Covid-19

No segundo trabalho, os pesquisadores usaram técnicas de *web-scraping* para coletar publicações da conta oficial da Comissão Nacional de Saúde da China, "*Healthy China*", do site Sina Weibo. Para cada publicação, capturaram: conteúdo do texto, número de curtidas, número de republicações, e número de comentários, além de links para fotos ou vídeo caso também fossem publicados, para determinar o tipo de mídia. Um total de 1441 publicações foram obtidas, com 1411 sendo relacionadas a Covid-19 por checagem manual.

A análise deles era composta de: engajamento dos cidadãos por meio da mídia social do governo (*CEGSM*), composto dos números de curtidas, republicações e comentários; riqueza de mídia, relacionado ao tipo de mídia usado (texto, imagens ou vídeos); laço dialógico, o qual envolve formas de interação com usuários, por exemplo: uso de enquetes ou *tags*; tipo de conteúdo, que era um resultado de uma análise de sentimento feita no texto da publicação.

O estudo mostrou que a riqueza de mídia estava negativamente associada com *CEGSM*, mostrando que copus composto de texto traziam mais engajamento, pelo menos durante o período da Covid-19. O estudo confirma que um laço dialógico trás

uma *CEGSM* maior.

2.3. Conclusão e similaridades

Todas as pesquisas abordam o tópico de engajamento que é o principal foco desse artigo, em geral dentro do mesmo período de tempo, mas voltado a contextos diferentes. Ainda assim, a conclusão geral de todos aborda assunto de desinformação e sua proliferação e intensificação ao longo da pandemia.

As métricas utilizadas foram usadas em parte nesse projeto, como o número de curtidas e comentários já que o acesso a tais informações via a ferramenta de coleta escolhida era mais fácil. Outras como análise de sentimento não foram utilizadas devido aos dados serem links e não texto. Como mostrado acima, o texto da publicação afeta o engajamento de forma considerável, então para fins dessa pesquisa, iremos desconsiderar tal efeito, focando em principal em publicações que não tem texto, composta somente pelos links.

3. Métodos e materiais

A coleta de dados foi dividida em duas partes. Primeiramente, a definição dos links das notícias que seriam escolhidas. Para isso, procurou-se um site com uma linha de tempo com links de notícias para que pudessem ser extraídas. O site encontrado foi o LARHUD IBICT, o qual contém uma linha do tempo da Covid-19 com diversos links disponibilizados pelo tempo de publicação. Importante relatar que o site não tinha notícias no mês de janeiro de 2020, então ele foi excluído da pesquisa.

O HTML foi extraído manualmente do site e depois foi processado por um script em Python 3.6 para extrair os links em conjunto com os domínios e a contagem de links de notícias por domínio. Essa contagem foi ordenada e os quatro domínios com maior número de links foram: *g1.globo.com*, com 344 links; *noticias.uol.com.br*, com 215; *agenciabrasil.ebc.com.br*, com 185; e *www.correiobraziliense.com.br*, com 96. Esses seriam os escolhidos, tendo os seus links uma distribuição uniforme por mês. Com os links e os domínios em mão, criou-se um script que pegaria aleatoriamente dois links por mês, almejando expressar entre os domínios.

Segundamente, a coleta das publicações foi feita via o site CrowdTangle, uma plataforma online dedicada à pesquisa que possibilita a pesquisadores coletar publicações nas grandes mídias sociais (Facebook, Instagram, Reddit e Twitter). Infelizmente, durante a coleta, notou-se que alguns links não tinham publicações encontradas, então foram substituídos por outros aleatoriamente. Assim, a distribuição que era uniforme por mês deixou de ser, apesar de ainda envolver os mesmos domínios. Os arquivos das coletas eram planilhas em formato csv, cada coluna representando uma feature e cada linha representando uma instância do conjunto, as quais eram as publicações propriamente ditas. Foram feitas 43 coletas, já que em fevereiro um dos veículos de notícia (Correio Braziliense) não tinha uma notícia publicada na linha do tempo. Já a quantidade de publicações variou dependendo da notícia e da empresa. Algumas tiveram um número extremamente pequeno de interações porém foram mantidas mesmo assim.

Após a coleta, foi realizado um pré-processamento, removendo colunas

irrelevantes para a anaĺise (como patrocinador, etc), mantendo somente aquelas que continham informações com relação a interação que a página teria (número de seguidores, curtidas, comentários, etc). Além disso, para valores utilizados nas métricas, como número de seguidores ou número de curtidas, se houvessem valores faltantes, ou eles eram substituídos um pelo outro (número de curtidas ia para número de seguidores ou vice-versa) ou a linha era removida.

Page Description	Page Created	Likes at Posting	Followers at Posting	Post Created	...	Message	Link	Final Link	Image Text	Link Text	Description
Fazer do Brasil a maior economia do planeta co...	2011-10-01 21:22:24	105288	103657.0	2020-03-24 15:20:41 BRT	...	Vamos ser realistas.... A inutil guerra as dro...	https://g1.globo.com/economia/noticia/2020/03/...	NaN	NaN	Coronavírus: com SP e RJ a partir desta terça,...	Quarentena obrigatória permite o funcionamento...
NaN	NaN	14682	NaN	2020-03-24 11:17:27 BRT	...	NÃO SUSPENDEM ATIVIDADES INDUSTRIAIS!!! Podem ...	https://g1.globo.com/economia/noticia/2020/03/...	NaN	NaN	Coronavírus: com SP e RJ a partir desta terça,...	Quarentena obrigatória permite o funcionamento...
NaN	NaN	15798	NaN	2020-03-25 12:35:39 BRT	...	QUEM MAIS SE SACRIFICOU ATÉ AGORA FORAM COMERC...	https://g1.globo.com/economia/noticia/2020/03/...	NaN	NaN	Coronavírus: com SP e RJ a partir desta terça,...	Quarentena obrigatória permite o funcionamento...
✵Deputado Estadual de São Paulo ✵ Autor de 48 l...	2011-11-23 12:43:33	6262	120259.0	2020-03-24 11:53:05 BRT	...	BRASIL: TODAS AS CAPITAIS FECHAM COMÉRCIO PARA...	https://g1.globo.com/economia/noticia/2020/03/...	NaN	NaN	Coronavírus: com SP e RJ a partir desta terça,...	Quarentena obrigatória permite o funcionamento...
Se mantenha informado de notícias locais e do ...	2016-07-07 11:50:09	8679	8869.0	2020-03-24 11:49:15 BRT	...	Brasil. https://g1.globo.com/economia/noticia/...	https://g1.globo.com/economia/noticia/2020/03/...	NaN	NaN	Coronavírus: com SP e RJ a partir desta terça,...	Quarentena obrigatória permite o funcionamento...

Figura 3. Amostra do dataframe com as publicações usando um link específico.

Por último, foi feita a análise na linguagem Python 3.6 dos dados, usando as features do dataset e fazendo comparações entre eles mesmos. Para essa, foi utilizado em principal a plataforma Google Collaboratory junto com as mais famosas bibliotecas de análise de dados como: pandas e matplotlib, em suas versões mais atuais. As métricas de análise foram: total de interação; total de interação média; taxa de engajamento baseada no número de seguidores; taxa de engajamento baseada no número de curtidas da página; taxa de engajamento baseada na média de ambas; alcance baseado no número de compartilhamentos. Além disso, também foi criado uma nuvem de palavras para observarmos quais eram os termos mais usados ao longo do ano.

Tudo isso foi implementado utilizando instalações físicas do instituto de computação da UFRJ e seus laboratórios ou o computador de mesa pessoal do integrante da pesquisa.

4. Resultados

Os resultados das comparações podem ser observados nos gráficos. Com relação a interação, grande parte dela ocorreu no meio do ano, ao longo de maio, junho, julho e agosto. Contudo, não é incomum que esse tipo de desnível ocorra devido ao fato de um perfil com muita influência, como um político ou celebridade, divulgar uma notícia e, assim, aumentar e muito o número de interações dela.

Entre as interações totais e médias, as distribuições ficaram muito similares, com picos no meio do ano e menos interações no início e fim. Isso pode ser devido a amostra que foi pega, no entanto. Mesmo que aleatória, ela pode não mostrar a realidade geral

das interações nesses períodos.

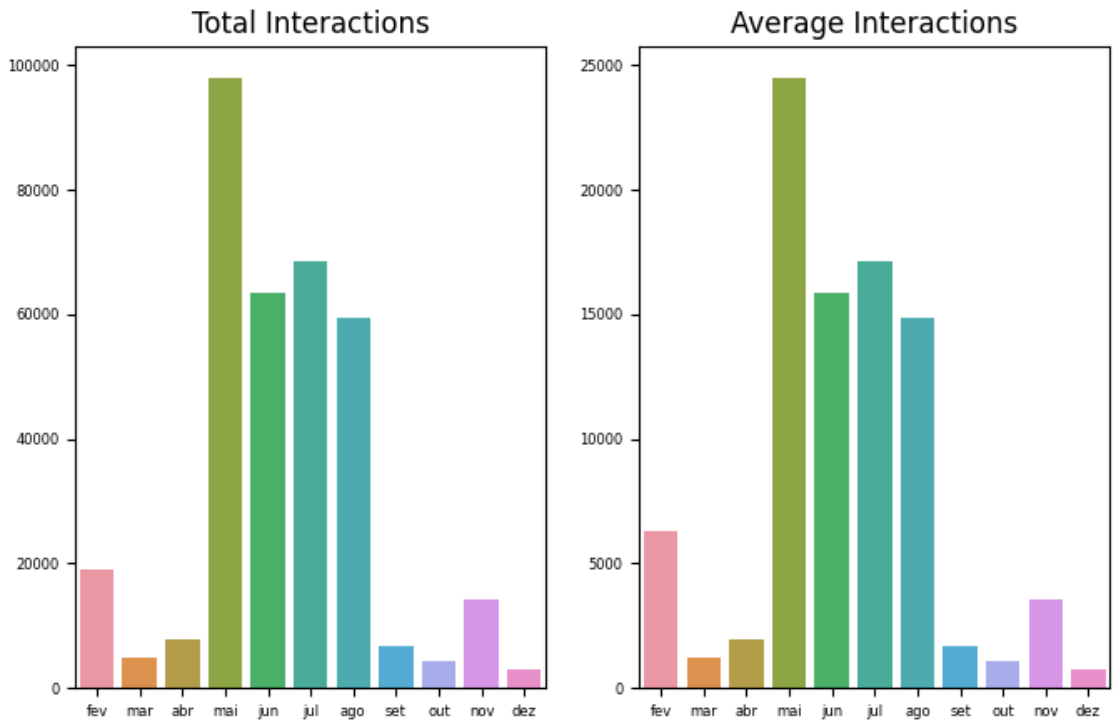

Figura 4. Gráfico do total de interações e das média de interações.

Já a taxa de engajamento é calculada baseado no número de total de interações da publicação dividido pelo número de seguidores, ou de curtidas, ou a média de ambos.

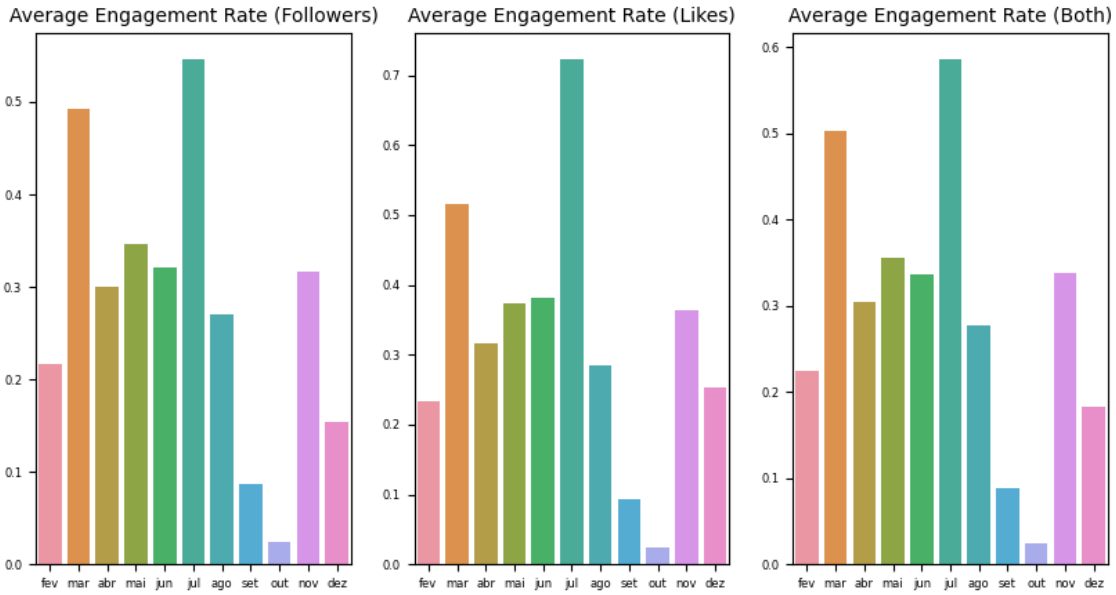

Figura 5. Gráfico das taxas de engajamento ao longo do ano de 2020.

A distribuição da taxa de engajamento mostra o fenômeno comentado mais acima, que existe uma grande possibilidade de uma página grande ter publicado uma notícia e assim atrair um grande número de interações. Pode-se observar que existe grande

engajamento nos meses de março e abril, meses iniciais da pandemia e isolamento social; e também em novembro, com as conversas sobre vacinação.

Outra forma de enxergar essas métricas é colocando suas distribuições por veículo de mídia.

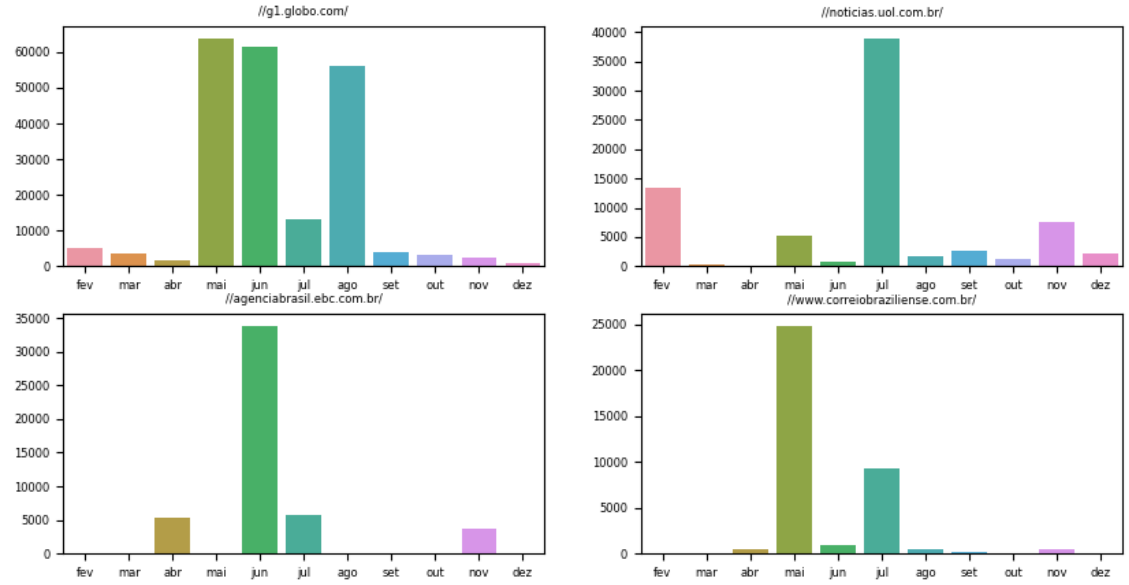

Figura 6. Gráfico do total de interações por veículo de mídia do ano de 2020.

Infelizmente, esses gráficos não nos dizem muita coisa além do que já foi afirmado acima. Pode-se perceber uma maior quantidade de interações nos domínios UOL e G1 pois são empresas bem maiores que Agência Brasil e Correio Braziliense. Apesar de parecer que não há interações em alguns meses, a realidade é que o número foi tão baixo que não aparece no gráfico. As únicas vezes que links foram substituídos da amostra foi quando nenhuma publicação foi encontrada.

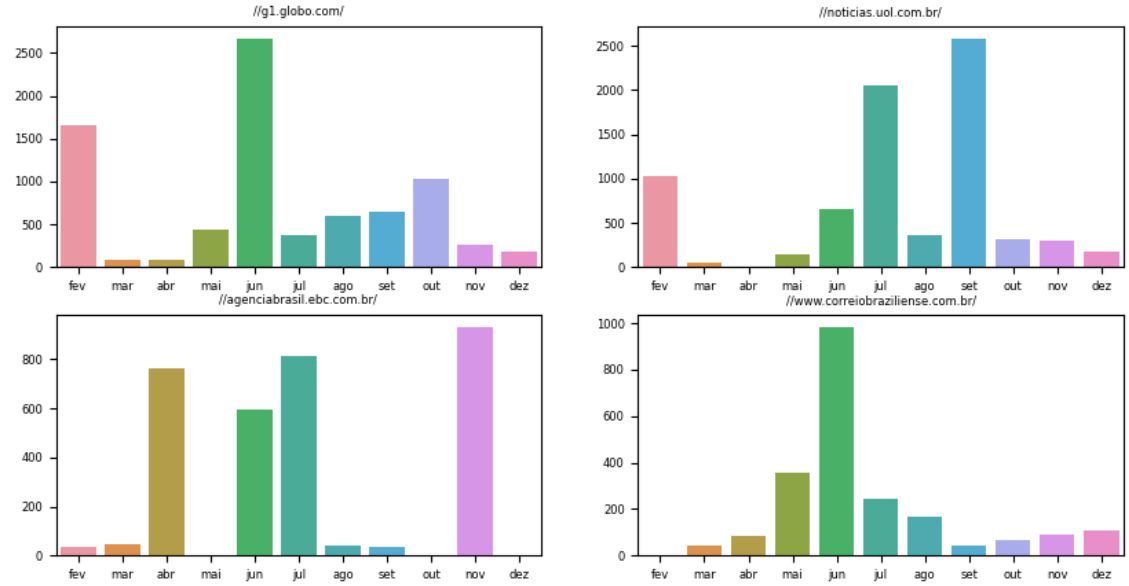

Figura 7. Gráfico do total de interações em média por veículo de imprensa do ano de

2020.

Novamente pode-se observar que quando feito o total de interações em média por publicação há uma suavização da distribuição, mostrando que existe entre os links, notícias que atraíram um perfil muito grande, ou até mesmo foram publicados pelas páginas das próprias empresas, que em geral tem uma boa quantidade de seguidores.

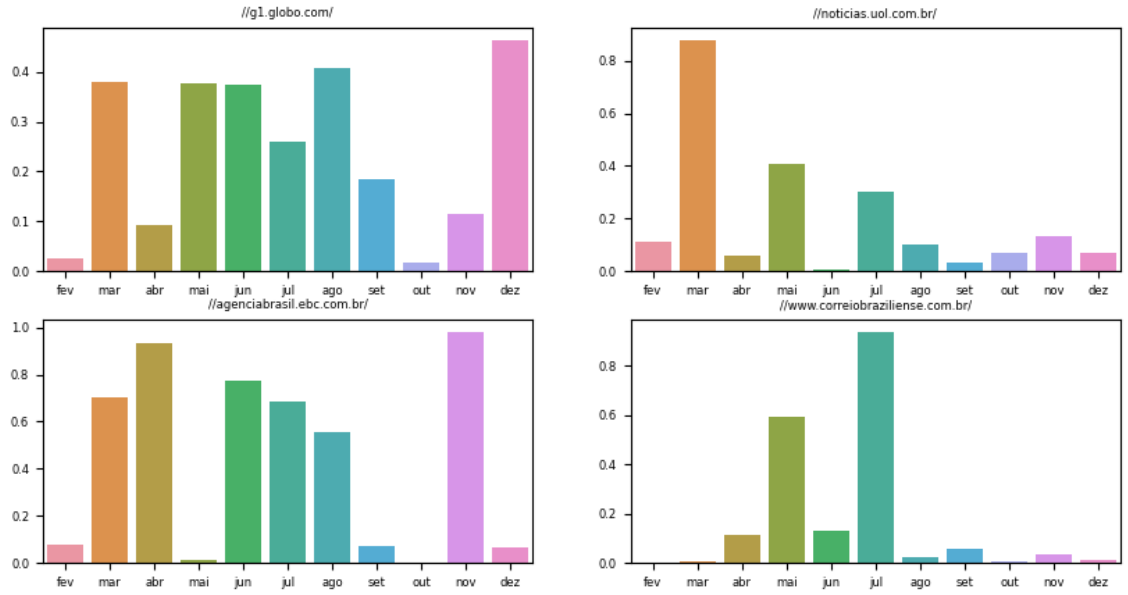

Figura 8. Gráfico da taxa de engajamento por seguidores por veículo de imprensa do ano de 2020.

Curiosamente, aqui observamos que a taxa de engajamento dos veículos de mídia menores é mais alta, em média, do que as dos maiores. O fato é que, pelo cálculo que a métrica de taxa de engajamento é realizado, é muito mais difícil uma página com muito renome conseguir um alto engajamento já que para isso ela precisa movimentar um número ainda maior de interações.

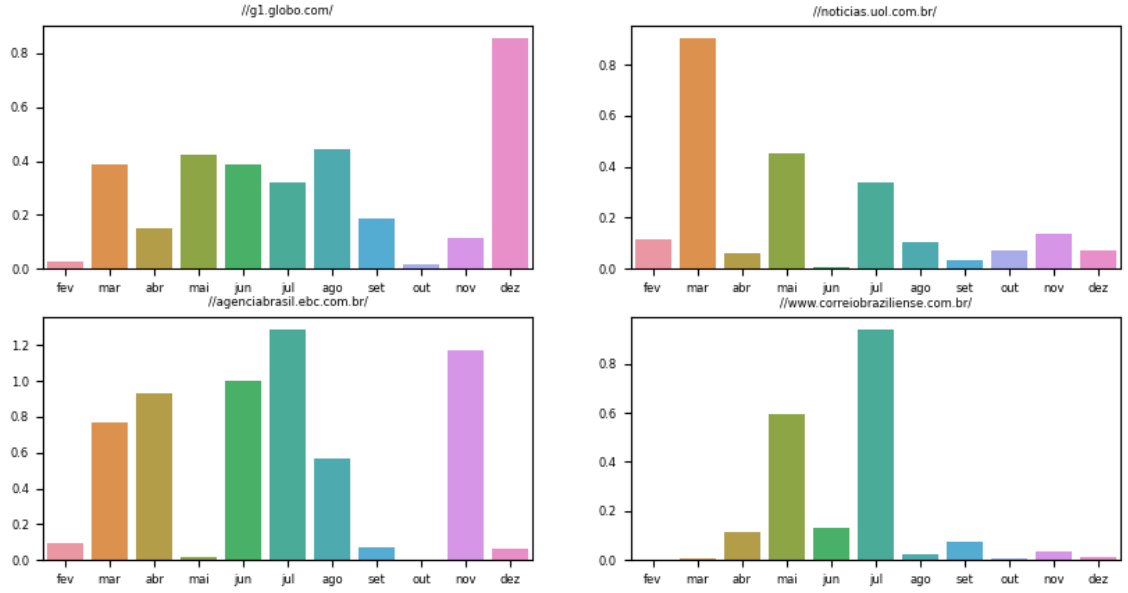

Figura 9. Gráfico da taxa de engajamento por curtidas por veículo de imprensa do ano de 2020.

Como é observado na imagem acima e na próxima imagem também, a distribuição não altera consideravelmente independentemente se se usa o número de curtidas, ou o número de seguidores, ou a média de ambos, o que facilita na hora do cálculo, pois como pode ver, fazer para cada um desses é consideravelmente mais custoso em questão de tempo. Além disso, caso falte uma informação ou outra, dá para se assumir um valor próximo mesmo assim.

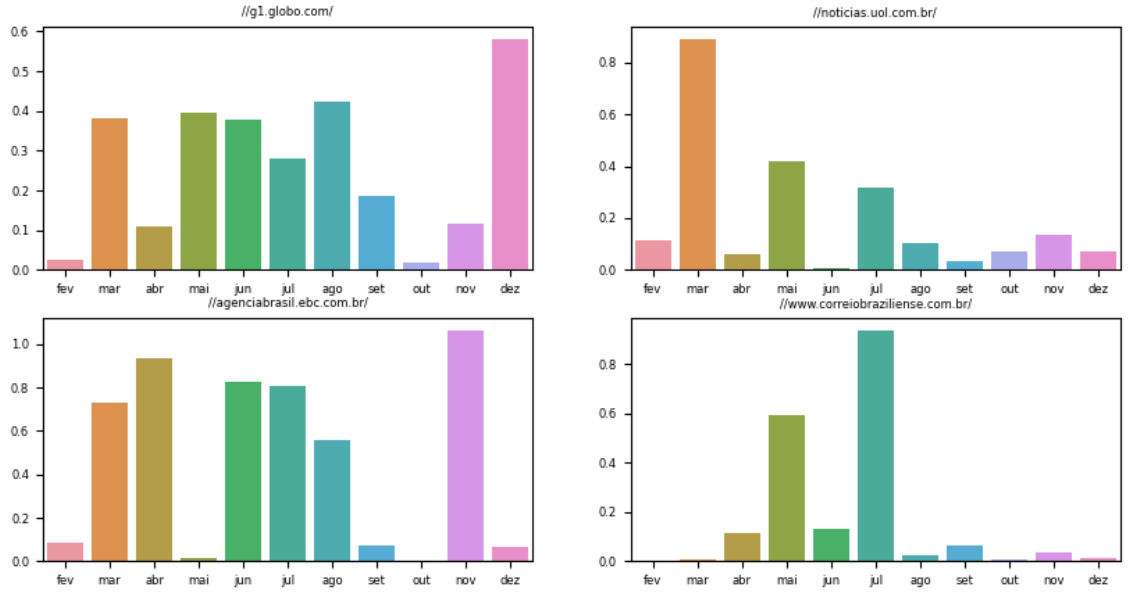

Figura 10. Gráfico da taxa de engajamento por ambos por veículo de imprensa do ano de 2020.

Por último, iremos mostrar o alcance por veículos de imprensa. Nos trabalhos relacionados, eles usam outras métricas para calcular alcance, como menções, por exemplo. O alcance aqui corresponde ao número de compartilhamentos.

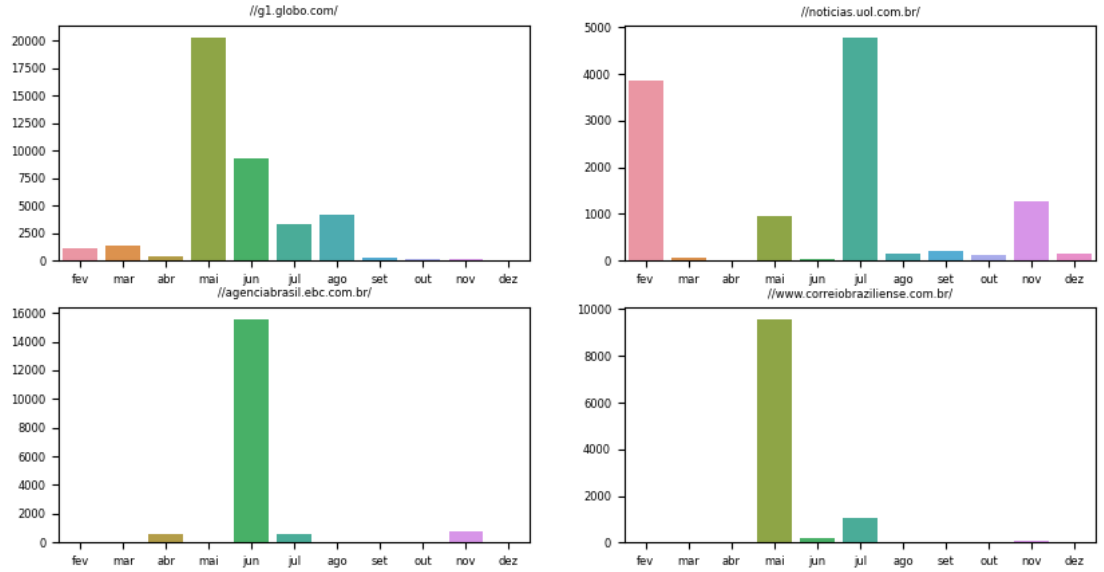

Figura 11. Gráfico da taxa de alcance por veículo de imprensa do ano de 2020.

Não há informação nova a ser retirada aqui. Nas próximas imagens, será possível ver as notícias que tem pico de interação e taxa de engajamento terminando com uma nuvem de palavras dos títulos e da descrição.

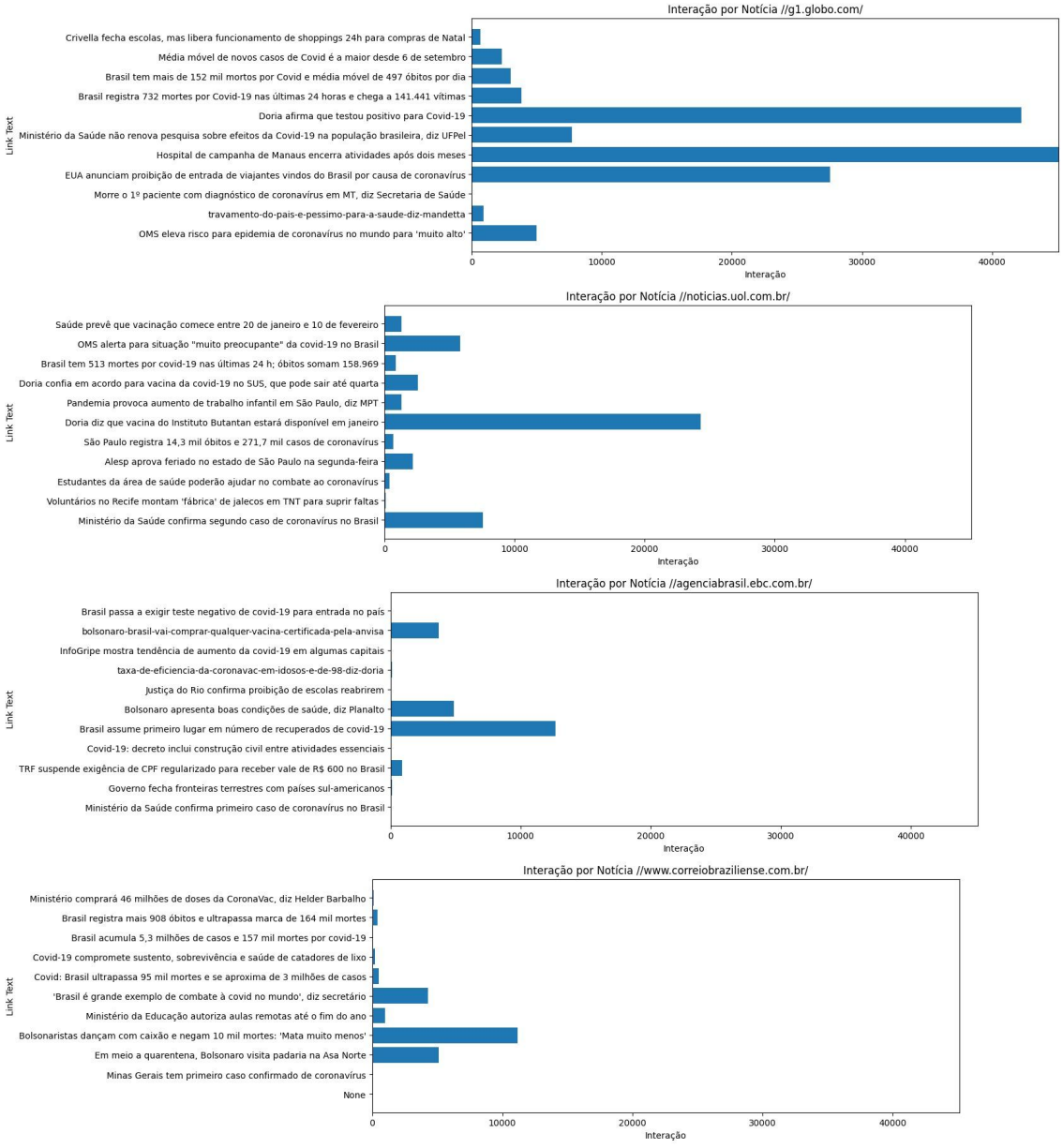

Figura 12. Gráficos das notícias por interação com veículos de mídia em 2020.

Na imagem acima, temos em exibição todas as notícias utilizadas na amostra e suas quantidade de interações. Outro fenômeno que pode ser observado é que notícias que envolvem políticos em geral tem alto engajamento ou número de interações. Em segundo plano, notícias que envolvem aspectos do Brasil como um todo, positivas ou negativas, também tendem a receber mais atenção.

Observe que em primeiro lugar tem uma notícia que refere um governador de São Paulo pelo nome, e, assim como no gráfico do G1, tem alta interação. Em segundo e

terceiro, notícias que envolvem o Brasil e mencionam o nome no título.

Ainda assim, é importante observar a discrepância entre o primeiro lugar e o segundo e o terceiro. Pode haver mais do que somente uma referência “preferida” como fator de engajamento, como comentado anteriormente.

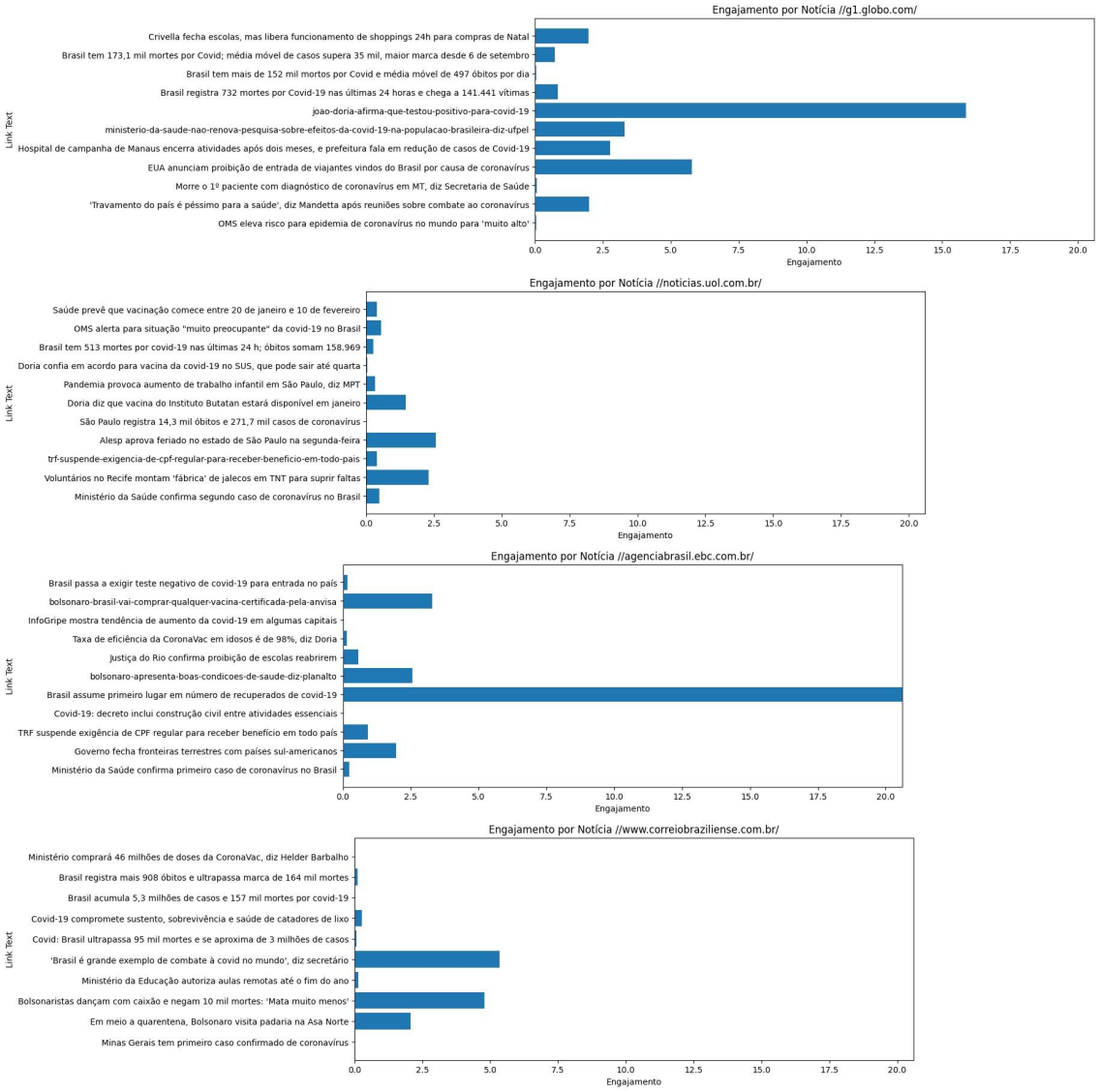

Figura 13. Gráficos das notícias por engajamento com veículos de mídia em 2020.

No engajamento, presencia-se o mesmo acontecimento dos gráficos de interação, onde o maior veículo de mídia tem maior engajamento em média e as notícias com mais engajamento são aquelas que se referem ao Brasil ou a políticos.

Por último, foi criado uma nuvem de palavras para ver quais as palavras mais utilizadas tanto nos títulos das notícias quanto nas descrições das publicações com as notícias.

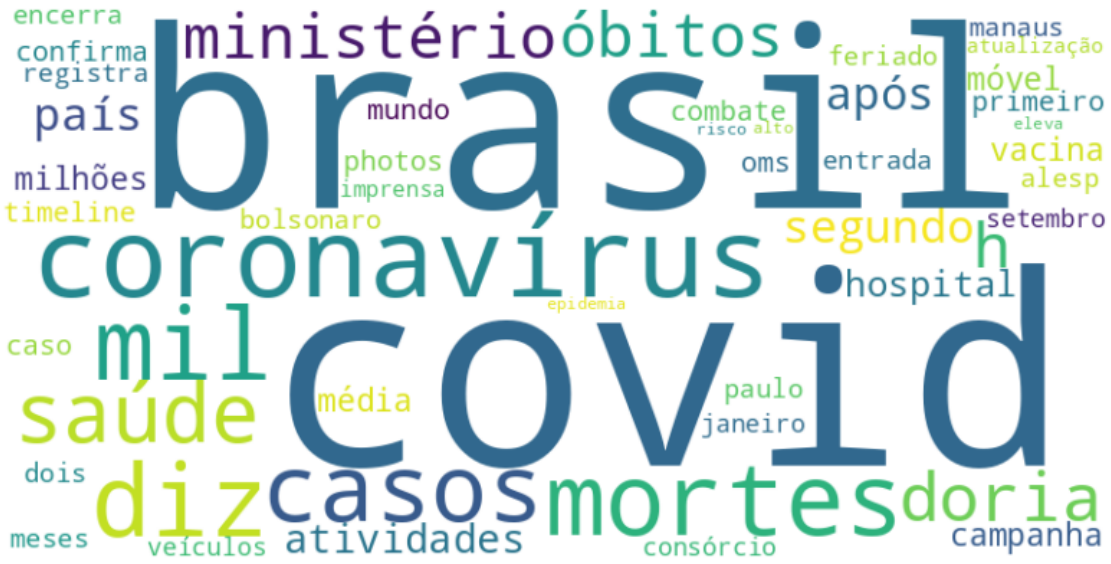

Figura 14. Nuvem de palavras de todos os títulos das notícias.

As palavras com maiores frequências foram ‘Brasil’, ‘covid’ e ‘coronavírus’, porém, note que muitas outras palavras nos títulos tem um aspecto negativo, voltado à doença, como: ‘casos’, ‘óbitos’, ‘mortes’. Uma tática bem comum para tentar atrair mais engajamento foi focar na doença em si, uma tática que existe a anos.

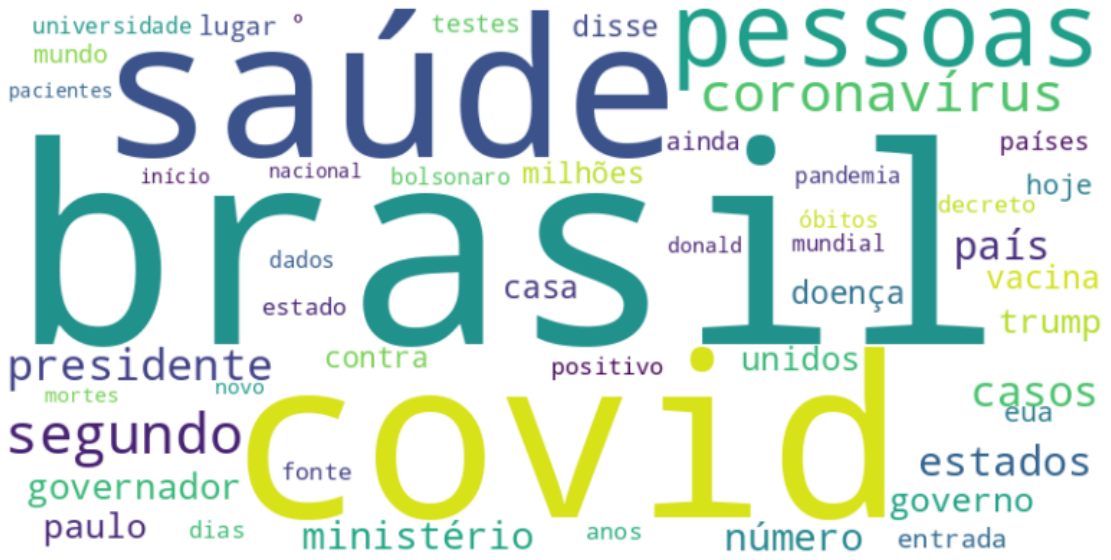

Figura 15. Nuvem de palavras de todas as descrições que acompanharam as notícias.

As descrições das publicações envolveram as palavras ‘covid’, ‘coronavírus’ e ‘Brasil’ também, mas, nota-se mudança no foco negativo que existia nos títulos das notícias, tendo ‘saúde’ em terceiro lugar e em todas as outras colocações as palavras eram mais diversificadas, tendo termos como ‘pessoas’, ‘presidente’, ‘segundo’.

5. Avaliação e Discussão

O primeiro ponto deve ser levantado são os possíveis viés que essa amostra

apresenta. Não só são 43 notícias aleatórias que podem dar uma impressão irreal do tipo de notícia que realmente traz engajamento, como elas próprias já tem um viés de escolha pois foram retiradas de uma seleção disponibilizada na linha do tempo. Quem colocou lá, já tinha em mente uma narrativa, ou tem viés próprio sobre quais tipos de notícias gostaria de exibir, mesmo que inconscientemente. Uma forma de diminuir a influência que esses possíveis viés causam seria aumentar o número de links e pegar outros veículos de imprensa menores, saindo dessa pequena bolha formada por esses quatro.

Outro problema que ocorreu foi o fato de não ter conseguido fazer alguma correlação entre as páginas que divulgaram as notícias e as próprias. Uma das ideias era fazer uma rede ligando o gênero da página a cada um dos veículos de mídia e verificar como ficaria esse grafo, porém o número de páginas com gênero indefinido era extremamente grande e não permitiria fazer uma rede representativa dessa amostra.

6. Conclusão

As principais conclusões removidas do experimento, que já não foram comentadas são: o renome do veículo de imprensa sempre é um fator; a taxa de engajamento muitas vezes pode ser associada a uma publicação "fora da curva"; os títulos de notícias em geral tinham palavras associadas com aspectos mais negativos; e as notícias que mais tinham engajamento e interação tinham a ver com governantes ou aspectos do brasil como um todo (em sua maioria, positivos).

Para trabalhos futuros, o ideal seria conseguir uma amostra maior de sites, envolvendo mais veículos de imprensa. Se possível uma forma automatizada de baixar os arquivos com as publicações, investir na correlação entre as páginas e as notícias que elas publicavam e utilizar o '*Overperforming Score*' como métrica para algum tipo de análise, já que ele vem nos arquivos do CrowdTangle e já está processado.

7. Agradecimentos

Direciono meus agradecimentos a professora Jonice pelo tempo, orientação e paciência fora das aulas, ajudando e estabelecer um plano de ação e guiando o que fazer quando perdido.

Ao Silas Filho, que mesmo à época de apresentação do seu projeto de doutorado, tirou tempo de sua agenda apertada para dar dicas de caminhos a prosseguir, com possíveis bibliotecas e ferramentas a explorar.

Ao fim, gostaria de agradecer aos meus pais, que têm a paciência do próprio Buddha para esperar eu cursar a minha graduação.

8. Referências

"Casos novos de COVID-19 por Semana Epidemiológica de notificação". https://covid.saude.gov.br/

Coelho, B., (2019) Os diferentes tipos de pesquisa científica. "Qual se aplica melhor a você?". https://blog.mettzer.com/tipos-de-pesquisa/

"Number of Facebook users in the United States from 2018 to 2027". https://www.statista.com/statistics/408971/number-of-us-facebook-users/

Qiang, C., et al. (2020) "Unpacking the black box: How to promote citizen engagement through government social media during the COVID-19 crisis". https://www.sciencedirect.com/science/article/pii/S0747563220301333

Massarani, L. et al. (2020) "O debate sobre vacinas em redes sociais: uma análise exploratória dos links com maior engajamento". https://www.scielo.br/j/csp/a/wg8Tn5R77L5v7YKJGPNcRYk/?lang=pt

Massarani, L. et al. (2021) "Infodemia, desinformação e vacinas: a circulação de conteúdos em redes sociais antes e depois da COVID-19" https://www.arca.fiocruz.br/handle/icict/51878