A Characterization of Portuguese Tweets Regarding the Covid-19 Pandemic

Pedro V. Brum, Matheus C. Teixeira, Renato Miranda, Renato Vimieiro, Wagner Meira Jr e Gisele L. Pappa

Universidade Federal de Minas Gerais, Brazil
{pedrobrum, matheus.candido, renato.miranda, rvimieiro, meira, glpappa}@dcc.ufmg.br

Abstract. Twitter has been one of the main sources of information and discussion during the COVID-19 pandemics. This paper characterizes a set of more than 56 million tweets written in Portuguese and collected over a period of 70 days. Our analysis includes the volume of messages, text of tweets, location of tweets, the main elements of tweets (e.g. hashtags and URLs) and the user profiles, including gender. The analyses showed the most discussed topics in the period were quarantine, hydroxychloroquine, agglomeration and social distance, and that the discussions were centered in political issues (e.g., most common hashtags include “fechadocombolsonaro" and “forabolsonaro").

CCS Concepts: • **Applied computing**;

Keywords: coronavirus, Twitter, social media, epidemics, public health, pandemic, covid-19

1. INTRODUÇÃO

Redes sociais *online* são importantes fontes de informação e discussão. A partir delas, os usuários difundem suas opiniões e se informam sobre diversos tópicos, tais como política, educação, saúde e esporte. Através de conexões de amizade e relações seguidor/seguido, as redes sociais permitem que seus participantes se comuniquem de forma efetiva, o que contribui para a discussão de tópicos variados [Bail et al. 2018; Guerra et al. 2013]. Em particular, o Twitter, com aproximadamente 316 mil usuários ativos por mês [Ahmed et al. 2019], é uma das maiores redes sociais da atualidade.

Dado o alcance mundial dessas plataformas, elas têm sido usadas como uma das principais fontes de informação sobre a pandemia do *coronavirus* (Covid-2019) desde os primeiros casos de infecção pelo Sars-CoV-2 reportados em Wuhan, na China, em dezembro de 2019. Os usuários, ao mesmo tempo que consomem as informações sobre a pandemia nas redes, expressam suas opiniões e posições em relação ao assunto. Eles repercutem medidas preventivas e possíveis tratamentos da doença, como a adoção de medidas de isolamento social e o uso de medicamentos como a hidroxicloroquina, além dos posicionamentos políticos e ações tomadas pelos governantes nos diferentes níveis, federal, estadual e municipal. Dessa forma, a análise do perfil dos usuários engajados nessas discussões, bem como das próprias discussões acerca da doença pode auxiliar na formulação de campanhas publicitárias para promoção de medidas preventivas, detecção de notícias falsas, entre várias outras aplicações. O principal objetivo deste trabalho é apresentar uma visão geral dos dados do Twitter durante a pandemia de Covid-19, em particular, de tweets em português.

A análise dos dados de redes sociais tem se tornado um assunto de extremo interesse na comunidade médica e se mostra necessária diante da atual pandemia de Covid-19 [Ahmed et al. 2019; Jiménez-Zafra et al. 2019]. Em particular, é interessante analisar os dados da rede do Twitter porque eles

podem ser facilmente coletados (desde que sejam públicos) e apresentam um volume suficiente para análises como as propostas nesse artigo. A grande maioria dos estudos relacionados à análise de redes sociais procuram apresentar uma visão geral de uma determinada comunidade em relação a um tópico especifico, tais como detecção de *bots* [Ferrara 2020], propagação de notícias falsas [Kouzy et al. 2020] ou polaridade política [Bail et al. 2018]. No contexto da pandemia de Covid-19, os estudos de análise de redes sociais apresentam informações demográficas e correlações de dados virtuais com dados reais [Menni et al. 2020].

Nessa direção, este trabalho apresenta uma análise dos dados do Twitter referentes ao COVID-19 publicados em língua portuguesa do período de 23 de abril a 02 de julho de 2020 (70 dias). O principal objetivo deste trabalho é realizar a caracterização dos dados e dos usuários interagindo com mensagens no contexto da pandemia de Covid-19, comparando os dados e indicadores ao longo do tempo. Em um segundo momento, esses dados padrão ser explorados para as aplicações já mencionadas, incluindo o combate a desinformação.

2. REVISÃO DE LITERATURA

Diversos trabalhos já foram propostos no contexto do estudo de dados demográficos referentes a pandemia de Covid-19. Dentre eles, [Dowd et al. 2020] consideraram dados demográficos de quatro países: Itália, Brasil, Nigéria e Coréia do Sul. A partir da análise dos dados demográficos, os autores conseguiram explicar como a idade da população pode influenciar na taxa de mortalidade por Covid-19 em um país e destacam a importância da disponibilidade de dados para o estabelecimento de medidas de combate ao coronavírus. Foi mostrado que a taxa de mortalidade por Covid-19 tende a ser maior em países onde a idade média da população é alta. Seguindo essa mesma linha de trabalho, [Nepomuceno et al. 2020] apresentaram como taxas de doenças crônicas podem afetar a taxa de mortalidade por Covid-19, mesmo em países onde a população é considerada jovem.

Já em [Chen et al. 2020] os autores apresentaram como foi feita uma coleta de dados do Twitter durante o começo da pandemia do novo coronavírus. A coleta foi realizada entre 21 de janeiro e 20 de março. A partir da coleta, uma base de dados foi construída considerando vários idiomas, tais como inglês, espanhol, japonês, tailandês, italiano, turco, indonésio e português. Além da coleta, os autores fizeram uma análise dos tweets, incluindo como a frequência de palavras relacionadas à pandemia (Por exemplo: "covid", "coronavirus") variaram ao longo do tempo. Também foi apresentado a variação do número de tweets em espanhol, italiano e japonês no período considerado.

Neste trabalho, apenas tweets em português são considerados. Porém, a base de dados construída possui 56.052.459 tweets, diferente da base de dados apresentada em [Chen et al. 2020], que possui apenas 3.451.196 tweets em português. Além disso, este trabalho apresenta uma análise mais abrangente dos dados do Twitter, levando em consideração tanto o conteúdo dos tweets como os usuários e a localização geográfica.

Já em [Kouzy et al. 2020], os autores analisaram o grau de desinformação que está sendo propagada no Twitter sobre a pandemia do novo coronavírus. Para realizar essa análise, foi necessário realizar uma coleta de tweets utilizando 14 hashtags e palavras-chave relacionadas à pademia de Covid-19, de forma similar ao que foi realizado neste trabalho.

3. COLETA DOS DADOS

A coleta dos dados foi realizada através da API do Twitter, utilizando as seguintes palavras-chave: “corona”, “covid”, “coronavirus”, “covid19”, “quarentena”, “hidroxicloroquina”, “cloroquina”, “confinamento”, “distanciamento social”, “aglomeração”, “aglomerações”, “sars” e “covid-19”. Assim, os tweets em português postados entre 23 de abril e 2 de julho de 2020 (70 dias) que possuíam pelo menos uma dessas palavras-chaves foram coletados. As *hashtags* foram consideradas da mesma forma. A base

Semana	Início	Fim	# Tweets	# Retweets
1	2020-04-23	2020-04-29	5.980.371	3.534.051
2	2020-04-30	2020-05-06	6.460.809	4.029.613
3	2020-05-07	2020-05-13	7.655.671	5.130.924
4	2020-05-14	2020-05-20	9.036.192	6.163.876
5	2020-05-21	2020-05-27	5.834.647	3.682.143
6	2020-05-28	2020-06-03	4.596.508	3.011.631
7	2020-06-04	2020-06-10	5.151.689	3.388.870
8	2020-06-11	2020-06-17	4.135.384	2.638.993
9	2020-06-18	2020-06-24	3.997.070	2.503.657
10	2020-06-25	2020-07-02	3.204.118	1.960.271

Table I: Número de tweets e retweets por semana.

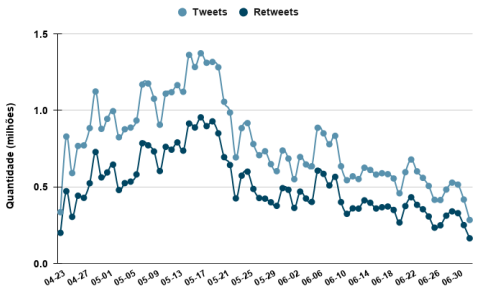


Fig. 1: Número de tweets e retweets por dia.

possui 56.052.459 tweets, sendo 36.044.029 (64,30%) destes retweets.

4. CARACTERIZAÇÃO DOS DADOS

A análise dos dados foi divida em 5 partes: (i) análise do volume de tweets ao longo do tempo; (ii) análise do texto dos tweets; (iii) análise da localização dos usuários que criaram os tweets; (iv) análise do número de tweets por usuário; e (v) análise das entidades (hashtags, urls e menções) presentes nos tweets.

4.1 Volume de Tweets

Os dados foram divididos em 10 partes, cada uma correspondente a um intervalo de tempo de 7 dias (uma semana) para a realização de uma análise temporal e do volume de tweets publicados. A Tabela I apresenta os intervalos de tempo considerados para a análise, além do número de tweets e retweets para cada um deles. É importante ressaltar que o conjunto de retweets está incluso no conjunto de tweets. A Figura 1 apresenta a variação do número de tweets e retweets por dia. É possível observar um aumento gradativo do número de tweets da semana 1 até a 4, onde ocorre o maior pico, que vai do dia 14 até 20 de maio. Mais especificamente, os números de tweets e retweets atingem o pico no dia 17 de maio, com 1.373.868 tweets. Dessa quantidade, 955.107 são retweets (69,52% dos tweets). Isso pode ser explicado pelo grande crescimento do número de casos de coronavírus[1], pela instabilidade política no Brasil provocada por demissões de ministros[2] e pelo debate sobre a eficácia da cloroquina como tratamento da Covid-19 [3]. A partir da semana 7, o número de tweets segue uma tendência de queda até a semana 10. Pela análise do gráfico, não é possível notar uma diferença considerável entre as variações do número de tweets e retweets ao longo do tempo. Foi possível observar ainda que no dia com a menor quantidade de tweets foram criados 284.156 tweets e 164.418 retweets (dia 2 de julho), e que a média do número de tweets por dia no período considerado foi igual a 789.471,25.

4.2 Análise Textual

Para análise textual dos tweets, realizamos primeiramente um pré-processamento do texto, onde foram removidos de cada tweet caracteres não alfanuméricos, acentos, *stop words* (por exemplo: artigos definidos e indefinidos, e preposições), sinais de pontuação, URLs, menções a usuários e *hashtags*. Além disso, todas as letras dos tweets foram colocadas em minúsculo e as ocorrências da expressão "covid 19" foram substituídas por "covid19".

[1]BBC Brasil: https://www.bbc.com/portuguese/brasil-52732620
[2]CNN Brasil: https://www.cnnbrasil.com.br/politica/2020/05/15/nelson-teich-pede-demissao-do-ministerio-da-saude
[3]Folha de São Paulo: https://folha.com/sxaz3y10

(a) Tweets (b) Retweets

Fig. 2: Nuvem de palavras dos tweets coletados.

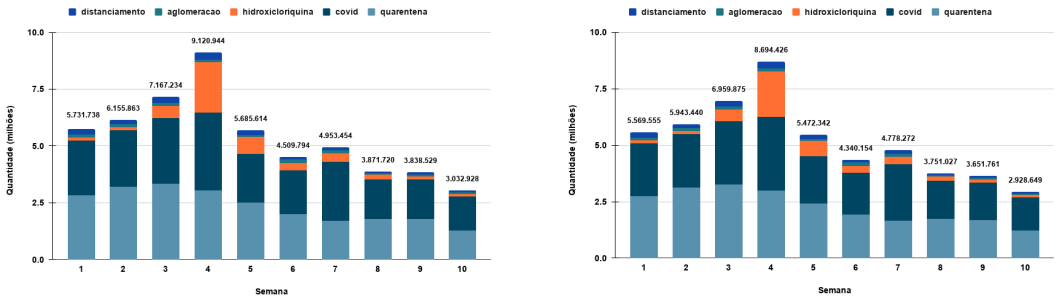
A partir do conjunto de 56.052.459 tweets foi possível identificar 1.161.283 tokens únicos. A fim de extrair os tokens mais frequentes, todas as palavras-chave (aquelas utilizadas como semente da coleta) foram removidas. Foi possível verificar que os vinte tokens mais frequentes, sem a remoção de duplicatas nos tweets, são (do mais frequente para o menos frequente): "brasil", "dia", "casa", "gente", "pessoas", "ser", "virus", "bolsonaro", "agora", "mortes", "1", "fazer", "saude", "sobre", "ter", "mundo", "pandemia", "casos", "tudo" e "acabar". A Figura 2 apresenta a nuvem de palavras para o conjunto de tweets do período de 23 de abril até 2 de julho. É possível observar que algumas das palavras mais relevantes do conjunto de tweets e retweets são "gente", "vou", "vai", "casa", "dia", "brasil", "virus", "mortes" e "bolsonaro". A partir da observação do conjunto de palavras-chave, foi possível identificar, cinco tópicos relacionados à pandemia de Covid-19: "covid", "quarentena", "hidroxicloroquina", "distanciamento" e "aglomeração". A Tabela II apresenta a lista de palavras relacionadas à cada tópico, bem como a frequência absoluta de cada um deles e o número de tweets em que eles ocorrem no período de 23 de abril até 2 de julho.

Tópico	**Palavras que definem o tópico**	**Frequência**	**Tweets**
Quarentena	quarentena	23.476.764	22.811.575
Covid	covid, covid19, covid-19, corona, coronavirus	22.854.936	22.007.481
Aglomeração	aglomeração, aglomerações	1.123.013	1.092.621
Distanciamento	distanciamento, confinamento, isolamento	1.706.198	1.622.295
Hidroxicloroquina	hidroxicloroquina, cloroquina	4.906.907	4.555.529

Table II: Tópicos relevantes discutidos no Twitter durante a pandemia do coronavírus.

A Figura 3 apresenta a proporção de ocorrência de cada tópico por semana em relação à frequência absoluta e número de tweets. A análise dos gráficos permite identificar um crescimento das palavras relacionadas ao tópico "Hidroxicloroquina" na semana 4, que vai do dia 14 até 21 de maio. Esse crescimento possivelmente ocorreu devido a liberação do uso da cloroquina para tratamento da Covid-19 pelo Minitério da Saúde nessa semana. É possível notar que os tópicos mais frequentes em todas as semanas são "Covid" e "Quarentena". Nota-se ainda uma redução da frequência absoluta dos tópicos relacionados a pandemia a partir da semana 5, assim como foi possível observar na Figura 1. Além disso, pode-se verificar uma redução na frequência do tópico "Distanciamento", o que reflete o relaxamento da população em relação às medidas de segurança propostas por órgãos de saúde para o combate ao coronavírus[4].

[4]VEJA: https://veja.abril.com.br/brasil/por-que-o-brasil-se-tornou-campeao-mundial-da-desordem-na-quarentena/



(a) Frequência absoluta

(b) Número de tweets

Fig. 3: Frequência de tópicos por semana.

4.3 Distribuição geográfica das postagens

Esta seção apresenta uma análise da distribuição geográfica das postagens contidas nos nossos dados. Para realizar esta análise, a localização do usuário foi extraída a partir de duas fontes: (i) localização geográfica exata dos tweets criados pelo usuário, dada pelos dados de latitudes e longitudes dos usuários que realizaram as postagens; e (ii) a localização autodeclarada pelo usuário, e que aparece em seu perfil do Twitter. Se para um determinado usuário foi possível extrair a informação sobre sua localização a partir de ambas as fontes, apenas a localização extraída a partir da primeira foi considerada. Assim, a localização de um determinado usuário foi definida como sendo o local onde o usuário criou o tweet ou o local autodeclarado por ele.

É importante ressaltar que boa parte dos tweets não possui a informação sobre o local de criação, uma vez que é necessária a concessão de permissão dos usuários ao Twitter para que suas localizações sejam registradas. Além disso, muitos dos locais autodeclarados estão incorretos ou insuficientes. Isso ocorre porque a localização autodeclarada corresponde a um campo aberto, em que os usuários podem inserir qualquer texto. Esses fatores dificultam a tarefa de identificação das localizações dos usuários.

Dos 56.052.459 tweets da base de dados, 10.559.339 (18,84%) foram criados no Brasil ou por usuários brasileiros. A partir do conjunto de tweets criados por usuários no Brasil, foi possível contar o número de tweets em cada uma das 27 unidades federativas. A Figura 4 apresenta a distribuição de tweets por estado. A partir da análise do gráfico, é possível observar que São Paulo (14,14%) é o estado com o maior número de tweets, seguido por Minas Gerais (11,14%), Rio de Grande do Sul (7,10%), Rio de Janeiro (7,02%) e Paraná (6,67%). Além disso, é possível verificar que os quatro estados com os menores números de tweets pertencem à região Norte e que três das cinco cidades com os maiores números de tweets pertencem à região Sudeste. A partir da Figura 2b é possível analisar o número tweets em cada uma das cinco regiões do Brasil (Norte, Nordeste, Centro-Oeste, Sul e Sudeste), ordenados de forma crescente, bem como o tamanho da população em cada uma delas [5]. É possível observar que a região com os maiores números de tweets é a Sudeste (35,52%). Além disso, é possível notar que apesar de ter uma população menor em relação à região Norte, a região Centro-Oeste possui um número maior de tweets.

4.4 Análise de URLs, hashtags e menções a usuários

No conjunto de 56.052.459 tweets considerado para análise, foi possível identificar 3.158.607 (5,64%) tweets com pelo menos uma *hashtag*, 4.406.366 (7,86%) tweets com pelo menos uma URL e 41.324.636 (73,73%) tweets como pelo menos uma menção a um usuário. Foi possível observar que a variação no número de tweets com pelo menos uma menção segue o mesmo padrão que o número de retweets.

[5]IBGE (população estimada em 2019) : `https://www.ibge.gov.br/cidades-e-estados`

Região	Tweets	%	População	Tweets/ População
Sudeste	3.750.105	35,52	88.371.433	0.042
Nordeste	2.696.776	25,54	57.071.654	0.047
Sul	2.009.567	19,03	29.975.984	0.067
Centro-Oeste	1.381.525	13,08	16.297.074	0.085
Norte	721.151	6,82	18.430.980	0.039

Table III: Número de tweets por região.

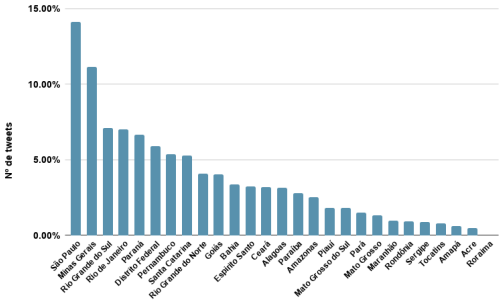
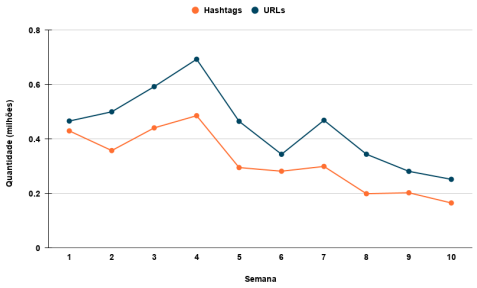
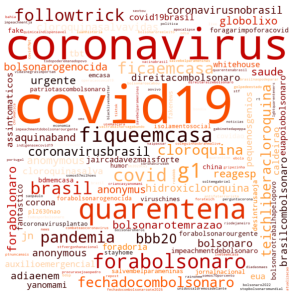


Fig. 4: Número de tweets por estado.

Isso é natural, uma vez que todo retweet possui uma menção ao autor do tweet original. A Figura 5a apresenta o número de tweets com *hashtags* e URLs por semana. A partir da análise do gráfico, é possível observar que, assim como os tweets e retweets, os números de tweets com *hashtags* e URLs atingem o seu pico na semana 4. Pode-se notar ainda um crescimento de tweets com URLs da semana 6 para a semana 7, o que não ocorreu com o número de tweets com *hashtags*.



(a)



(b)

Fig. 5: (a) Número de tweets com *hashtags* e URLs (b) Nuvem de palavras para hashtags.

No conjunto de tweets foi possível identificar 313.082 *hashtags* únicas. A partir disso, foi possível observar a ocorrência de *hashtags* muito similares, tais como "COVID19", "Covid19" e "covid19". Para contornar esse problema, foi necessário a aplicação do pré-processamento sobre o conjunto de *hashtags*, de forma a padronizar o texto. Dessa forma, os *hashtags* "COVID19", "Covid19" e "covid19" são todos mapeados para "covid19". Depois do pré-processamento foi possível identificar 247.916 *hashtags* únicos. A Figura 5b apresenta a nuvem de palavras para o conjunto de *hashtags*. As vinte *hashtags* mais frequentes são (em ordem decrescente em número de tweets): “covid19”, “coronavirus”, “quarentena”, “g1”, “fiqueemcasa”, “forabolsonaro”, “covid”, “brasil”, “followtrick”, “ficaemcasa”, “pandemia”, “cloroquina”, “fechadocombolsonaro”, “bbb20”, “teichliberacloroquina”, “coronavirusbrasil”, “forabolonaro”, “saude”, “globolixo” e “anomymous”.

4.5 Perfil dos usuários

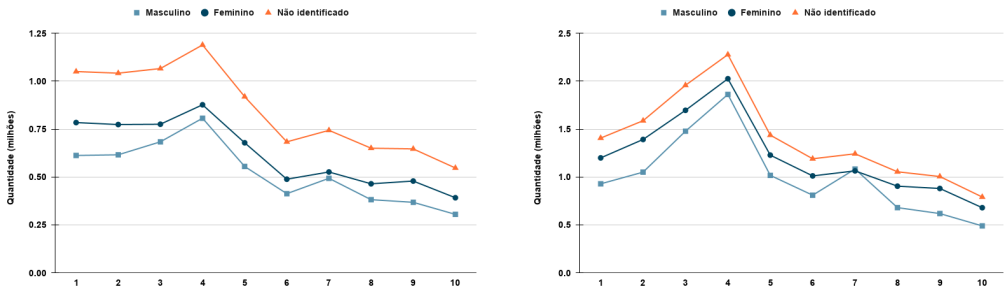
Esta seção apresenta uma análise do conjunto de usuários da base de dados. Ou seja, usuários que postaram ao menos um tweet no período entre 23 de abril e 2 de julho, em português, utilizando pelo menos uma das palavras-chave utilizadas para coleta. No conjunto de 56.052.459 tweets, foi possível identificar 5.232.337 usuários. Desse conjunto de usuários, 3.660.391 realizaram pelo menos um tweet original (usuário que criou o tweet também é o autor) e 3.772.343 realizaram pelo menos um retweet.

Foi possível observar que os usuários realizaram em média 10,71 tweets e que mais da metade dos usuários realizaram pelo menos 3 tweets. Além disso, foi possível verificar que o usuário que mais criou tweets realizou 285.520 postagens (usuário "*quarentena_bot*") no período entre 23 de abril e 2 de julho.

Alguns dos usuários que mais criaram tweets no período são usuários jornalísticos, ou seja, contas que representam indivíduos ou organizações que publicam conteúdos jornalísticos (Por exemplo: JornalDestak e SICNoticias). Esses usuários estão interessados apenas na difusão de notícias e não na opinião de indivíduos específicos. Também é possível que o conjunto de usuários que mais criaram tweets contenha usuários *spammers*, e que representam indivíduos ou organizações com interesse em difundir vírus, pornografia ou propagandas. Esse conjunto também pode conter, em menor quantidade, usuários legítimos, ou seja, que não são jornalísticos nem *spammers*.

Também foi possível inferir os gêneros dos usuários da base de dados. Essa tarefa foi realizada a partir de um dicionário de nomes [Miranda Filho et al. 2014], onde cada nome é classificado como feminino, masculino ou unisex (pode ser tanto masculino ou feminino). O nome do usuário pode estar disponível em duas fontes diferentes: o nome de usuário e o nome que aparece no perfil (i.e. pseudônimo), que o usuário escolheu como identificação. É importante ressaltar que o nome de usuário e o nome que aparece no perfil não necessariamente são nomes de pessoas. Ou seja, esses dois nomes podem corresponder apenas ao pseudônimo do usuário. Para a tarefa de classificação dos gêneros dos usuários foi dada prioridade ao nome de usuário. Assim, se não for possível inferir o gênero de um determinado usuário a partir de seu nome, o seu pseudônimo é utilizado. Seguindo essa estratégia, foi possível identificar o gênero de 61.07% dos usuários. Do conjunto de 3.195.193 usuários em que foi possível inferir o gênero, 44.94% foram classificados como sendo do gênero masculino e 55.06% como sendo do gênero feminino.

A Figura 6 apresenta a distribuição de tweets originais (de autoria do usuario que criou o tweet) e retweets por semana para os usuários do gênero masculino e feminino. A partir da análise dos gráficos, pode-se verificar que os usuários do gênero feminino realizaram mais tweets e retweets em todo o período especificado, exceto na semana 7, onde pode-se observar uma quantidade maior de retweets por parte dos usuários do gênero masculino. Pode-se observar também um aumento considerável de tweets originais e retweets para os usuários do gênero masculino entre a semana 6 e 7, o que corresponde ao intervalo de tempo entre os dias 28 de maio e 10 de junho.



(a) Tweets originais

(b) Retweets

Fig. 6: Número de tweets originais e retweets por semana para os usuários do gênero masculino e feminino.

5. CONCLUSÃO

Este trabalho apresentou uma análise de dados do Twitter do período entre 23 de abril e 2 de julho, considerando tanto o conteúdo dos tweets como os usuários e a localização geográfica. A análise foi

possível a partir da coleta de tweets que possuem hashtags ou palavras-chave relacionadas à pandemia de Covid-19.

Observou-se que o período com os maiores números de tweets corresponde ao intervalo entre os dias 14 e 20 de maio. Isso pode ser explicado sobretudo pela crise institucional no governo brasileiro com a segunda troca do Ministro da Saúde durante a pandemia. O fato da troca ter sido aparentemente motivada pela discordância entre o presidente e ministro em relação às políticas adotadas para o enfrentamento à doença fica evidente nos principais tópicos observados nos tweets. Entre os tópicos mais frequentes nos tweets, nós observamos "Quarentena", "Hidroxicloroquina", "Aglomeração" e "Distanciamento" que indicam também um debate ativo dos usuários em relação às medidas de enfrentamento da Covid-19. Em particular, o debate sobre o uso da hidroxicloroquina no tratamento da doença foi acentuado no período de 7 a 20 de maio, semanas 3 e 4 no nosso estudo. Nós observamos também que o debate foi mais intenso na região Sudeste, correspondendo a 35,53% das postagens durante o período de análise.

Esses tópicos revelam que o debate no país foi mais centrado nas políticas de saúde pública adotadas pelos governantes no enfrentamento à pandemia. Nesse sentido, planejamos estender este trabalho para caracterizar o posicionamento político dos usuários participantes, bem como os argumentos utilizados à favor e contra as medidas adotadas pelo Ministério da Saúde.

ACKNOWLEDGMENTS

Os autores agradecem a FAPEMIG, MPMG (projeto Capacidades Analíticas), CNPq, CAPES, MCTIC/RNP (51119), H2020 (777154), MASWeb, INCT-Cyber, ATMOSPHERE e ao IFMG - Campus Sabará pelo apoio financeiro.

REFERENCES

Ahmed, W., Bath, P. A., Sbaffi, L., and Demartini, G. Novel insights into views towards h1n1 during the 2009 pandemic: a thematic analysis of twitter data. *Health Information & Libraries Journal* 36 (1): 60–72, 2019.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Lee, J., Mann, M., Merhout, F., and Volfovsky, A. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115 (37): 9216–9221, 2018.

Chen, E., Lerman, K., and Ferrara, E. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance* 6 (2): e19273, 2020.

Dowd, J. B., Andriano, L., Brazel, D. M., Rotondi, V., Block, P., Ding, X., Liu, Y., and Mills, M. C. Demographic science aids in understanding the spread and fatality rates of covid-19. *Proceedings of the National Academy of Sciences* 117 (18): 9696–9698, 2020.

Ferrara, E. What types of covid-19 conspiracies are populated by twitter bots? *First Monday* 25 (6): 1–25, 2020.

Guerra, P. H. C., Meira Jr, W., Cardie, C., and Kleinberg, R. A measure of polarization on social media networks based on community boundaries. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2013.

Jiménez-Zafra, S. M., Martín-Valdivia, M. T., Molina-González, M. D., and Ureña-López, L. A. How do we talk about doctors and drugs? sentiment analysis in forums expressing opinions for medical domain. *Artificial intelligence in medicine* vol. 93, pp. 50–57, 2019.

Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M. B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E. W., and Baddour, K. Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter. *Cureus* 12 (3): e7275, 2020.

Menni, C., Valdes, A. M., Freidin, M. B., Sudre, C. H., Nguyen, L. H., Drew, D. A., Ganesh, S., Varsavsky, T., Cardoso, M. J., Moustafa, J. S. E.-S., et al. Real-time tracking of self-reported symptoms to predict potential covid-19. *Nature medicine* vol. 26, pp. 1037–1040, 2020.

Miranda Filho, R., Carvalho, A. I., and Pappa, G. L. Inferência de sexo e idade de usuários no twitter. In *Anais do III Brazilian Workshop on Social Network Analysis and Mining*. SBC, pp. 200–211, 2014.

Nepomuceno, M. R., Acosta, E., Alburez-Gutierrez, D., Aburto, J. M., Gagnon, A., and Turra, C. M. Besides population age structure, health and other demographic factors can contribute to understanding the covid-19 burden. *Proceedings of the National Academy of Sciences* 117 (25): 13881–13883, 2020.