# Mineração de Texto no *Twitter*: uma ferramenta auxiliar na detecção de epidemias

**Maria Denise Simões**[1], **Ana Régia de M. Neves**[1]

[1]Eixo de Informação e Comunicação – Instituto Federal de Brasília (IFB) – *campus* Brasília
Brasília – DF – Brasil

maria.simoes@estudante.ifb.edu.br, ana.neves@ifb.edu.br

***Abstract.*** *On March 11, 2020, the World Health Organization (WHO) has declared the novel coronavirus (Covid-19) outbreak a global pandemic. In Brazil, data from different official sources are not integrated, such situation has made it difficult for crosscheck the information and undermines the investigation of Covid-19 scenario in each region. The process of text mining on social media is being used for monitoring and to assist the early prediction of epidemics. In this context, the aim of this article is to adopt text mining approach to analyze the trend of Covid-19 in Brazil and testing the hypothesis that social media monitoring can help detect epidemics. The quantitative results confirm the hypothesis when compared to official data available from the government agencies.*

***Resumo.*** *Em 11 de março de 2020, a Organização Mundial da Saúde (OMS) declarou o estado de pandemia de Covid-19. No Brasil, os sistemas oficiais responsáveis pela visibilidade dos dados da pandemia não são integrados, o que dificulta o cruzamento de dados e prejudica a investigação do cenário da Covid-19 em cada região. O processo de mineração de texto em mídias sociais está sendo utilizado para deteção e monitoramento de doenças. Nesse contexto, o objetivo deste artigo é realizar uma mineração de texto na rede social Twitter, a fim de analisar a tendência epidemiológica de uma série temporal dos dados sobre Covid-19 no Brasil, testando a hipótese de que o monitoramento das mídias sociais pode auxiliar na detecção de epidemias. Os resultados quantitativos confirmam a hipótese levantada quando comparados aos dados dos agentes oficiais notificadores.*

## 1. Introdução

Os primeiros casos de Covid-19 foram relatados em Wuhan, na China, no final de 2019. Em seguida, devido à sua alta transmissibilidade, espalhou-se para diversos outros países, o que levou a Organização Mundial de Saúde a declarar, no dia 11 de março de 2020, estado de pandemia. De acordo com [Kraemer et al. 2020], em 2020, os casos de Covid-19 foram registrados em mais de 180 países.

No Brasil, existem dois sistemas oficiais do Ministério da Saúde que são os principais responsáveis pela visibilidade dos dados da pandemia de Covid-19, os quais são[1]:

[1]https://coronavirus.saude.gov.br/definicao-de-caso-e-notificacao

- o SIVEP-Gripe[2] (Sistema de Informação de Vigilância Epidemiológica da Gripe): além da notificação dos casos de Síndrome Gripal (SG), registra casos de Síndrome Respiratória Aguda Grave (SRAG) hospitalizados e óbitos (independente de hospitalização);
- o e-SUS Notifica[3]: foi desenvolvido para atender à alta demanda de notificações de casos suspeitos e confirmados de Covid-19;

A partir desses sistemas, cada estado consolida os seus dados e gera os boletins epidemiológicos diários para monitoramento e análise da situação. Porém, esses sistemas não estão integrados, o que dificulta o cruzamento de dados e prejudica a investigação do cenário da Covid-19 em cada território.

De acordo com um estudo realizado por pesquisadores do projeto Monitora Covid-19, lançado pelo Laboratório de Informação em Saúde (LIS), do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT), da Fiocruz [MonitoraCovid-19 2020], foram encontradas discrepâncias entre as datas, real e oficial, de ocorrência dos eventos (casos e óbitos) registrados nesses sistemas, o que também impacta nas análises de tendência da pandemia no país. Além disso, pode ocorrer a duplicidade de registros, já que a atualização dos resultados no e-SUS Notifica, por exemplo, não é automática.

O processo de mineração de texto aplicado às redes sociais, como o *Twitter*, pode ser utilizado na área da saúde pública para a detecção precoce de surtos epidêmicos, o rastreamento de doenças e para auxiliar na avaliação da conscientização da população sobre questões relativas à saúde [Jahanbin and Rahmanian 2020, Charles-Smith et al. 2015, Ginsberg et al. 2009].

Nesse contexto, o objetivo desse trabalho é realizar a mineração de texto na rede social *Twitter*, a fim de executar análise de série temporal dos dados sobre Covid-19 no Brasil, para confirmar a hipótese que o monitoramento do *microblogging* pode auxiliar na detecção e propagação de epidemias.

As demais Seções deste artigo estão dispostas como segue: a Seção 2 descreve a metodologia utilizada na revisão da literatura; a Seção 3 apresenta o processo de mineração de texto aplicado no *Twitter* para os dados sobre Covid-19; os resultados são apresentados e discutidos na Seção 4; e, por fim, a Seção 5 expõe as considerações finais e os trabalhos futuros para esta pesquisa.

## 2. Revisão Sistemática da Literatura

Este artigo é fundamentado de acordo com a revisão sistemática da literatura e direcionado pela seguinte questão: "É possível detectar sinais de epidemias por meio das postagens no *Twitter*, em tempo hábil para atuação das entidades públicas?".

As buscas foram baseadas no título e resumo dos trabalhos, e ocorreram no período de outubro a dezembro de 2020. Os critérios de inclusão e exclusão são descritos na Tabela 1. As fontes onde as buscas foram realizadas e o conjunto de *strings* utilizadas são apresentados na Tabela 2.

[2]https://sivepgripe.saude.gov.br/sivepgripe/login.html?0).
[3]https://notifica.saude.gov.br/onboard

**Tabela 1. Critérios definidos para a revisão sistemática**

| Critérios de inclusão | Critérios de exclusão |
|---|---|
| Aplicação da mineração de dados ou textos em mídias sociais | Predição de surtos epidêmicos sem o processo de mineração de dados ou textos |
| Predição de surtos epidêmicos por meio de mídias sociais | Manuscritos que não respeitaram o objetivo do estudo ou a pergunta norteadora |
| Mineração de texto no Twitter por técnicas de aprendizagem de máquina | Estudos em discordância com os critérios de inclusão |

**Tabela 2. Fontes e *Strings* de busca utilizadas**

| Fonte | *String* | Resultado |
|---|---|---|
| SciELO | (epdemic OR epidemia) AND (flu OR influenza) AND (covid19 OR covid-19 OR coronavirus) OR (mineração de dados) | 134 |
| SciELO | (epdemic OR epidemia) and (flu OR influenza OR covid19 OR covid-19 OR coronavirus OR dengue) and ( twitter OR redes sociais OR social network) | 13 |
| MEDLINE via PubMed | covid AND prediction AND data mining OR text mining | 29.629 |
| Ministério da Saúde | covid OR covid 19 AND predição AND aprendizagem de máquina | 32 |
| Periódicos CAPES | covid OR covid 19 AND twitter AND machine learning | 447 |

## 2.1. Resultados da Revisão Sistemática da Literatura

Após pesquisa nas bases eletrônicas da ScieELO, PubMed, Ministério da Saúde e Periódicos da Capes, 30.255 mil artigos foram identificados como possivelmente relevantes; 28.216 mil foram removidos porque estavam duplicados e/ou não apresentavam relação com o objetivo da pesquisa. Dos 2.039 mil artigos restantes, 2.018 mil foram excluídos após triagem baseada no resumo e título; e 14 artigos foram removidos por falta de informações. Finalmente, dos 7 artigos incluídos na síntese final, 2 foram identificados como trabalhos correlatos, pois aplicaram mineração de texto no *Twitter* para a predição de epidemias.

[Zhang et al. 2017] desenvolveram um modelo para estimar a incidência e progressão epidêmica da gripe sazonal por região, baseado em dados coletados no *Twitter*.

Em [Gomide et al. 2014], o objeto de estudo foi o mapeamento dos casos de dengue também utilizando dados do *Twitter* e, a partir de análises da correlação de *Spearman* e clusterização, identificou-se o mesmo comportamento das informações oficiais disponibilizadas pelo Ministério da Saúde.

## 3. Material e Métodos

Para executar a mineração de texto no Twitter, este trabalho se baseou no fluxo proposto por [Aranha 2007] que identifica cinco fases para esse processo, a saber:

1. Coleta: construção da base textual baseada na definição das strings e extração automática dos *tweets*;
2. Pré-processamento: definição da amostra e limpeza dos registros;
3. Indexação: não aplicado;
4. Mineração de dados: classificação manual dos *tweets*; e
5. Análise: interpretação e avaliação dos resultados.

Para a extração dos dados no *Twitter*, foi utilizada a biblioteca *Tweepy*[4] que permite a conexão com a API do *Twitter*. É possível recuperar *tweets* postados entre seis e nove dias anteriores a data da busca. O ambiente de execução do *crawler* foi o Google Colaboratory[5].

### 3.1. Coleta dos *tweets*

A coleta de *tweets* ocorreu no período de dezembro de 2020 a março de 2021. O idioma selecionado para a extração foi o português e os termos definidos para compor as *strings* de busca foram baseados nos trabalhos correlatos descritos na Seção 2.1, quais sejam:

- covid19 e sinônimos (corona, #coronavirusPlantao, covid-19, #covid19brasil, coronavirus, covid); e
- sintomas comuns (falta de ar, febre, e tosse).

A partir desses termos, foram compostas *strings* de busca com o objetivo de coletar o maior número de *tweets* e com a menor quantidade de repetições e/ou *retweets*. Por exemplo:

- String 1 ('covid19 OR corona OR #CoronavirusPlantao OR covid-19 OR #covid19brasil OR coronavirus OR covid OR covid 19') and ('falta de ar') and ('febre') and ('tosse');

Além disso, o último termo da *string* é o que mais aparece nos textos dos *tweets*, por isso, procurou-se intercalar os que definem sintomas (tosse, febre e falta de ar) a fim de que cada um fosse contemplado na última posição em todos os dias de coleta.

A etapa de coleta resultou em 59 mil registros de *tweets*.

### 3.2. Pré-processamento

Após a coleta, inicia-se a etapa de limpeza dos dados para prover uma representação estruturada do que foi recuperado. A qualidade da base influencia diretamente os resultados da análise.

Primeiro, foi feita a redução da dimensionalidade da base de acordo com a técnica de amostragem definida por [Arkin and Colton 1963], onde para uma população inicial de 50 mil observações e margem de erro desejada de no máximo 2%, é definida uma amostra ideal de 2.381 mil observações.

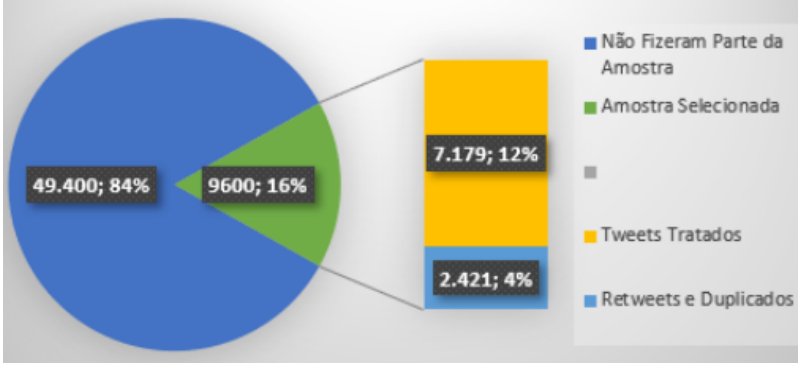



**Figura 1. Distribuição da base de dados coletada.**

Assim, dos 59 mil *tweets* coletados, 9.600 mil fizeram parte da amostra, dos quais 7.179 mil foram tratados para a limpeza de *retweets* e *tweets* duplicados. A Figura 1 apresenta a redução realizada.

Os *retweets* (RT) são postagens feitas originalmente por um usuário e compartilhados por outras pessoas no perfil. Neste artigo, considerou-se *tweets* duplicados as postagens feitas em sequência por um indivíduo descrevendo sintomas. Por exemplo, o primeiro *tweet* foi: "parece que peguei uma gripe"; depois, "sentindo mal, estou com febre". A limpeza de *retweets* e *tweets* duplicados é necessária porque, de acordo com [Li and Cardie 2013], esse tipo de informação pode inserir viés nos dados, prejudicando a construção e análise de tendências.

### 3.3. Mineração de dados

Para atingir o objetivo desta pesquisa é necessário classificar os *tweets* da amostra. É importante ressaltar que nem todos os *tweets* que contêm palavras-chave das *strings* de busca, definidas na Seção 3.1, indicam que o usuário está com Covid-19.

Assim, após a etapa de pré-processamento nos 7.179 mil *tweets*, foi realizada a classificação manual destes *tweets* de acordo com os rótulos abaixo listados:

- Críticas, Diversos (textos que não se identificam com o tema);
- Ansiedade;
- Ironia;
- Politização: crítica ao posicionamento ou pronunciamento de uma figura pública, em geral político;
- Outras Doenças;
- Notícias/Informações relacionadas;
- Covid-19; e
- Vacina.

Essas classes foram definidas com base nos termos utilizados nos trabalhos selecionado da revisão sistemática (Seção 2.1) e após avaliação dos tipos de postagens mais frequentes das que foram coletadas. A classe de *tweets* sobre Covid foi utilizada para a consolidação da base de dados e da construção de uma série temporal para análise.

[4]https://www.tweepy.org/
[5]https://research.google.com/colaboratory/intl/pt-BR/faq.html

## 4. Resultados e Discussão

A Figura 2 ilustra os 7.179 mil *tweets* após a classificação manual realizada. É possível verificar que dentre as postagens classificadas a mais frequente é referente à Covid, com 4.436 mil referências sobre o tema.

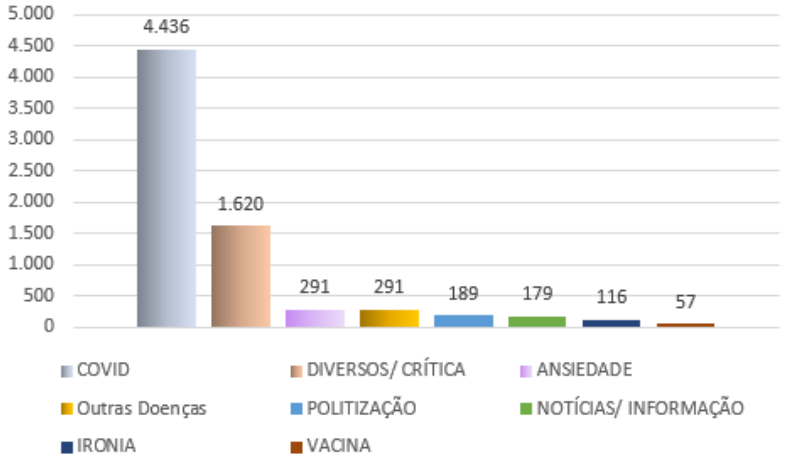



**Figura 2. Quantidade de *tweets* por categoria.**

A Figura 3 apresenta uma série temporal de 04 de fevereiro até o dia 07 de março de 2021, contabilizando 96 arquivos por dia (eixo *x*). A quantidade de *tweets* por arquivo é representada no eixo *y*.

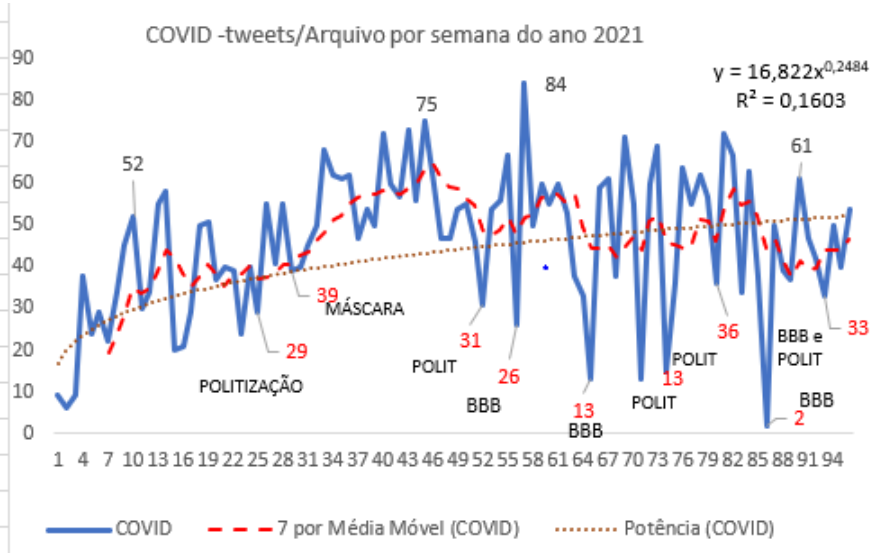



**Figura 3. Consolidação dos arquivos coletados.**

O gráfico apresentado na Figura 3 também contém a linha de tendência (em pontilhado) e a média móvel da categoria (em vermelho), a qual é calculada somando-se o número de casos de cada um dos sete dias anteriores e dividindo esse resultado por 7 (número de dias considerado).

A análise de tendência de série temporal, que é a medida estatística de quão próximos os dados estão da linha de regressão ajustada, apresenta regressão linear ($R^2$ =

0,1603) e tendência de aumento. A média móvel da série oscila entre os picos e vales e demonstra movimento ascendente ao final.

Os picos representam altos quantitativos de postagens sobre sintomas da Covid-19, testagens positivas, declarações de piora ou melhora desses sintomas do próprio usuário ou de alguém conhecido. Os vales são as ocorrências de *retweets* e seguidas postagens sobre: (i) a repercussão da notícia de que componentes do programa *Big Brother* Brasil (BBB) apresentarem sintomas da Covid-19; (ii) Politização (Polit); e (iii) debate sobre o uso da máscaras de proteção respiratória.

É importante ressaltar que alguns arquivos coletados registraram 100% de *retweets*, isso influenciou na queda da média móvel e da linha de tendência, podendo resultar em vieses na interpretação.

Tendo em vista que as epidemias, em geral, são tratadas por semana epidemiológica[6], buscou-se representar os dados também de modo similar para proceder com a comparação de tendências. Assim, a Figura 4 retrata uma série temporal da quinta até a nona semana de 2021 (eixo $x$), o que corresponde ao período de 04 de fevereiro de 2021 até o dia 07 de março de 2021; o eixo $y$ contém a quantidade de *tweets* classificados apenas sobre Covid-19.

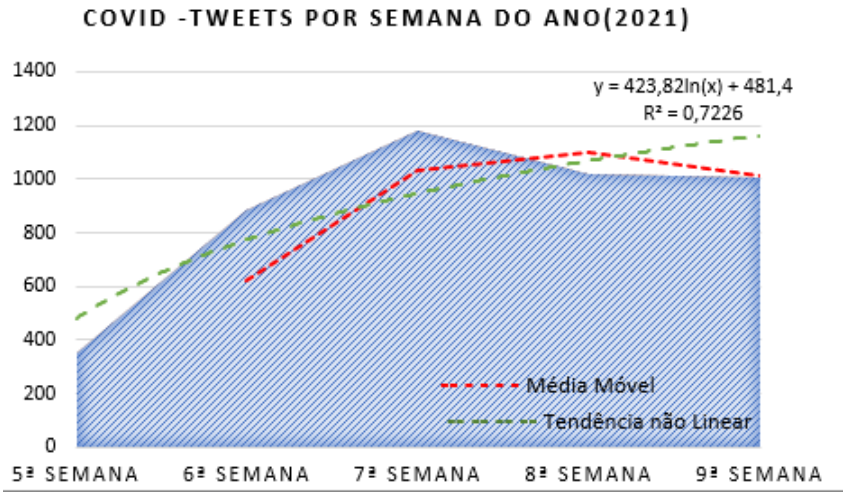



**Figura 4. *Tweets* sobre Covid-19 por semana, 2021**

O gráfico da Figura 4 também apresenta a análise de tendência da série temporal, no caso, não-linear, calculada por regressão logarítmica, com ($R^2 = 0{,}7226$) e tendência de crescimento positivo nas postagens sobre Covid-19. A média móvel (em vermelho) apesar de registrar tendência de queda, ainda apresenta altos números nas postagens.

### 4.1. Comparação de Tendência com os dados dos Órgãos Oficiais

Para verificar a hipótese de que o monitoramento das mídias sociais pode auxiliar na detecção de epidemias, foi realizada uma comparação entre a série temporal dos dados provenientes do *Twitter*, com os dados (i) do Ministério da Saúde (2021), (ii) da Fundação Oswaldo Cruz (FIOCRUZ), e (iii) da Organização Mundial da Saúde (OMS).

[6]http://saude.sp.gov.br/resources/cve-centro-de-vigilancia-epidemiologica/publicacoes/se_serie.pdf

A Figura 5 ilustra os dados consolidados pelo Ministério da Saúde, no período de 31/1 a 6/2/2021, para os casos notificados de Covid-19 ([de Vigilância em Saúde 2021]). O eixo $x$ descreve as semanas epidemiológicas e o eixo $y$ as quantidades de casos notificados. É possível verificar a tendência de aumento da média móvel.

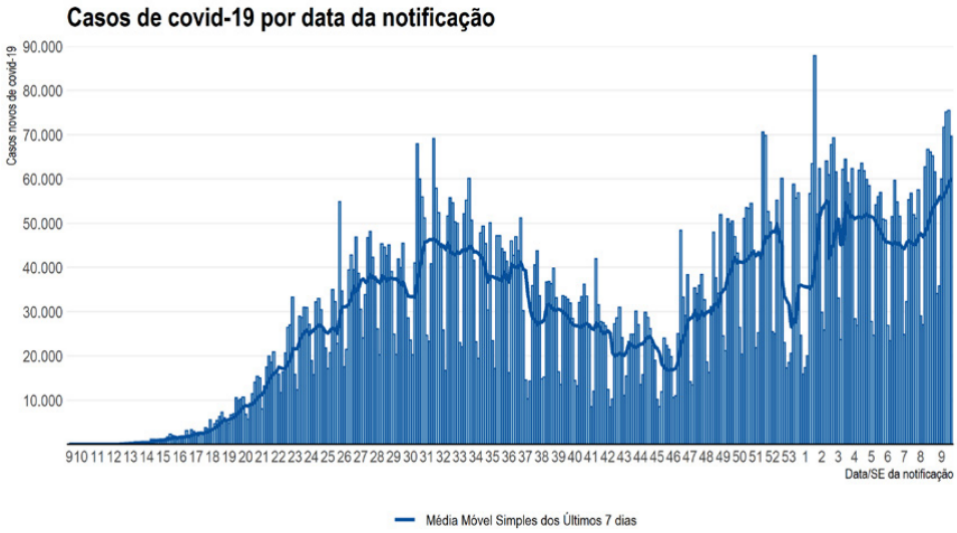
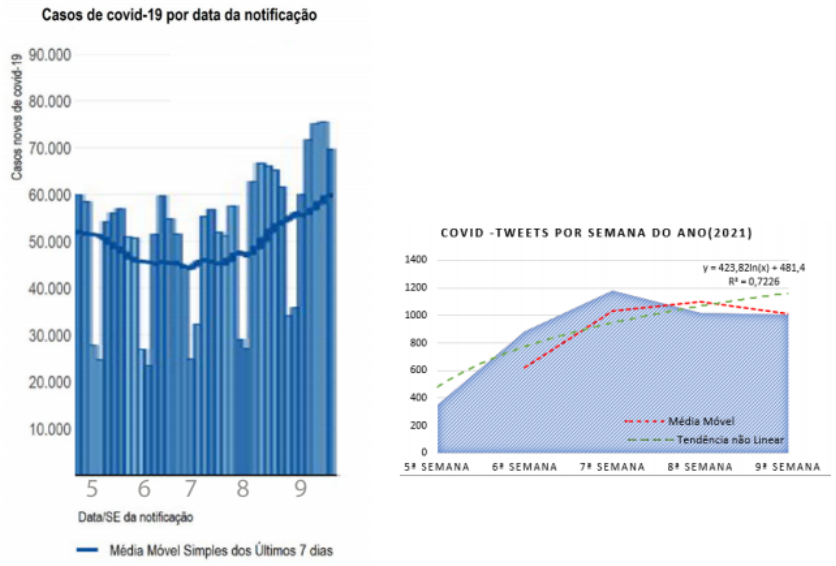



**Figura 5. Casos de Covid-19 notificados pelo Ministério da Saúde por semana epidemiológica.**

A Figura 6 apresenta a comparação entre o recorte realizado no gráfico da Figura 5 para o período da quinta até a nona semana de 2021 (Figura 6a), e a Figura 4 representando os dados coletados do *Twitter* no mesmo período (Figura 6b).




**(a)** Recorte do Ministério da Saúde

**(b)** *Tweets* sobre Covid-19 por semana, 2021

**Figura 6. Comparação de tendências entre o gráfico do Ministério da Saúde e os *Tweets* coletados sobre Covid-19**

Observa-se que a média móvel no recorte do Ministério da Saúde apresenta tendência de aumento (Figura 6a); enquanto que pelos *tweets* apresenta leve tendência de queda (Figura 6b).

Essa discrepância pode ser explicada devido ao comportamento dos usuários no *Twitter* ser influenciado por eventos externos, como o caso dos *retweets* sobre o BBB, o que compromete as análises, conduzindo a um viés na interpretação dessa medida estatística. Por isso, novos cálculos estatísticos foram realizados considerando o período da sexta a nona semana de 2021, conforme Figura 7.

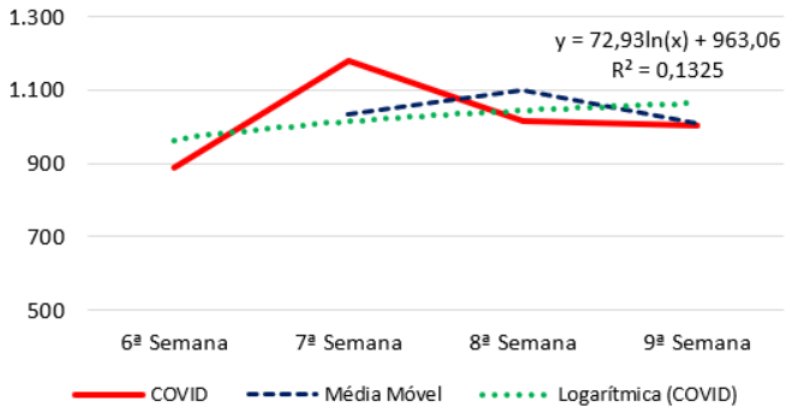



**Figura 7. *Tweets* da sexta a nona semana em 2021**

Para dados coletados a partir da sexta semana, a média móvel apresenta leve tendência de queda com regressão $R^2 = 0{,}01325$. No eixo $x$ estão dispostas as semanas do ano, a partir da sexta. O eixo $y$ apresenta as quantidades de *tweets*.

Os próximos dados comparados foram os disponibilizados pela FIOCRUZ[7], representados na Figura 8. No eixo $x$ estão registradas as semanas epidemiológicas e no eixo $y$, a quantidade de casos; bem como a média móvel (em azul) e as linhas de tendência.

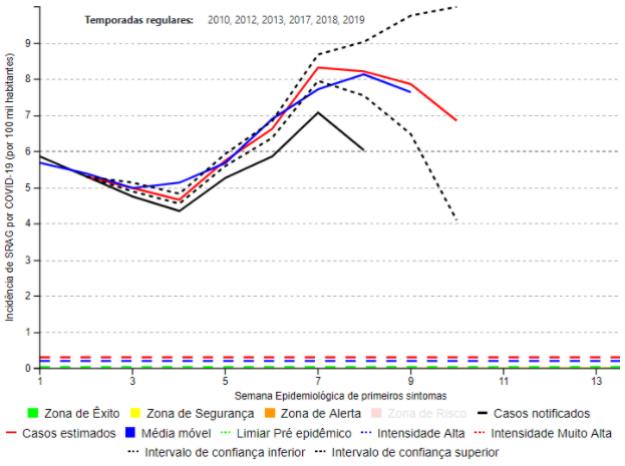
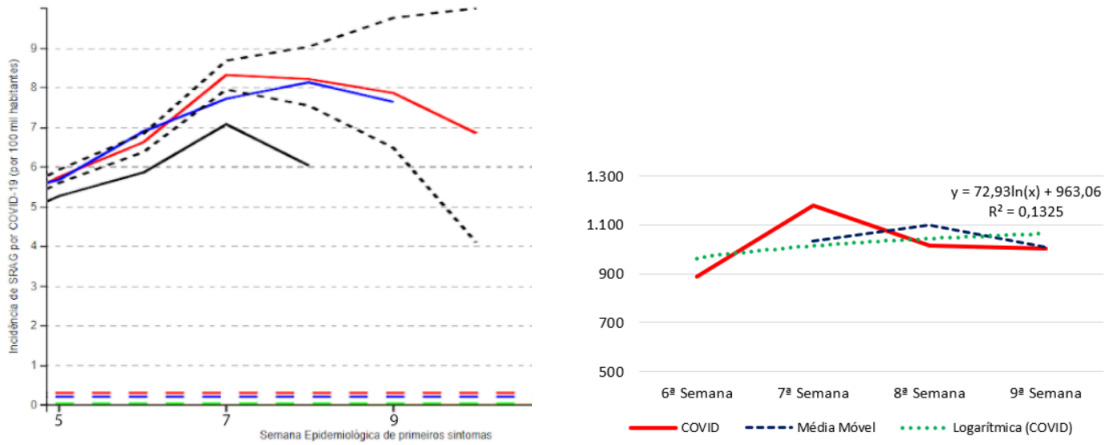



**Figura 8. Curva de incidência de SRAG por Covid-19, ano 2021**

[7]http://info.gripe.fiocruz.br/

O gráfico representado pela Figura 8 também foi recortado para indicar as semanas analisadas por este estudo, da sexta a nona semana de 2021. Lembrando que os dados referentes a quinta semana foram retirados devido à alta incidência de *retweets*. A Figura 9 apresenta a comparação entre o recorte da FIOCRUZ (Figura 9a) e os dados coletados do *Twitter* no mesmo período (Figura 9b).




**(a)** Recorte da FIOCRUZ

**(b)** Tweets da sexta a nona semana em 2021

**Figura 9. Comparação de tendências entre o gráfico da FIOCRUZ e os Tweets coletados sobre Covid-19**

A partir da comparação dos gráficos apresentados nas Figuras 9a e 9b, verifica-se que a curva da média móvel dos *tweets* acompanha a curva do gráfico da FIOCRUZ no mesmo período.

Por último, a Figura 10 apresenta o quantitativo de casos confirmados semanalmente no Brasil de acordo com o Painel do Coronavírus divulgado pela OMS[8].

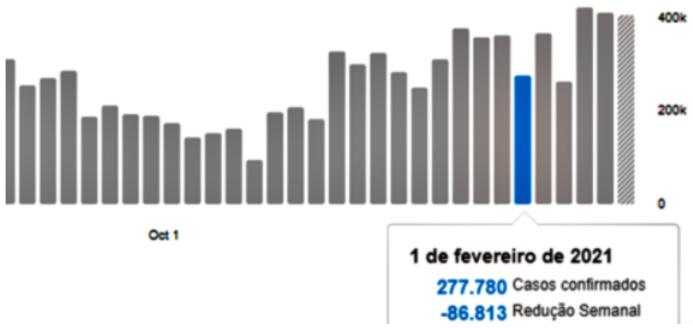
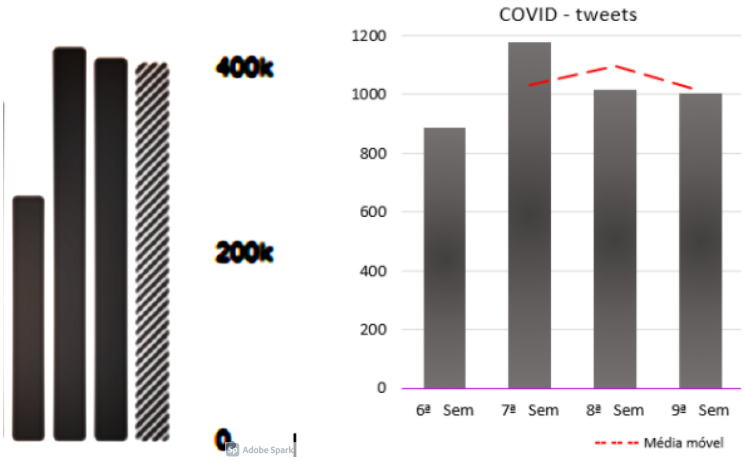



**Figura 10. Painel do Coronavírus da OMS, Brasil**

O gráfico da Figura 10 também foi recortado para representar as semanas analisadas por este estudo, da sexta a nona semana de 2021. A Figura 11 apresenta a comparação entre o recorte da OMS (Figura 11a) e os dados coletados do *Twitter* no mesmo período (Figura 11b).

[8]https://covid19.who.int/region/amro/country/br




**(a)** Recorte da OMS

**(b)** Gráfico de coluna dos *Tweets* da sexta a nona semana em 2021

**Figura 11. Comparação de tendência entre o gráfico da OMS e os *Tweets* coletados sobre Covid-19**

A partir da comparação entre as Figuras 11a e 11b, verifica-se a pequena tendência de queda em ambas as representações.

## 5. Considerações finais e Trabalhos futuros

O presente estudo foi fundamentado a partir de uma revisão sistemática da literatura, cuja metodologia trouxe a sustentação teórica para a pesquisa. Nessa fase, a partir de uma questão norteadora, foram sistematizados conhecimentos sobre mineração de texto no *Twitter*, deteção e predições de tendências de surtos epidêmicos, como Covid-19, dengue e *influenza*.

No Brasil, os sistemas oficiais de notificação não são integrados e ocorrem defasagens entre as datas de ocorrência dos eventos, o que impacta na tomada de decisão e a comunicação à população pelas autoridades. O que notabiliza a urgência na utilização de métodos alternativos para estimar a incidência de epidemias no território nacional.

A partir da mineração de texto, análise de série temporal e comparação com os dados disponibilizados pelos órgão oficiais, como Ministério da Saúde, FIOCRUZ e OMS, evidenciou-se a eficácia das postagens no *Twitter* como ferramenta auxiliar para a detecção de surtos epidêmicos.

É importante ressaltar que apesar do *Twitter* ser importante ferramenta para analisar tendências, seu monitoramento deve ser conduzido com cautela, pois constatou-se que eventos externos podem modificar o comportamento dos usuários e consequentemente a curva de tendência das variáveis analisadas, o que pode introduzir viés na interpretação dos resultados.

Como trabalhos futuros, indica-se o aumento dos registros coletados para uma série temporal maior, a inserção geolocalização para identificar os focos epidêmicos por território, a construção de um modelo preditivo para a classificação automática de *tweets*, permitindo uma análise de tendência mais eficaz e atuação mais célere na prevenção e

detecção de epidemias pelas entidades públicas.

## Referências


Alimohamadi, Y., Sepandi, M., Taghdir, M., and Hosamirudsari, H. (2020). Determine the most common clinical symptoms in covid-19 patients: a systematic review and meta-analysis. *Journal of Preventive Medicine and Hygiene*, 61(3):304–312.

Aranha, C. N. (2007). *Uma abordagem de préprocessamento automatico para mineração de textos em português: sob o enfoque da inteligência computacional*. PhD thesis, Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Arkin, H. and Colton, R. R. (1963). *Tables for statisticians*. Barnes and Noble.

Charles-Smith, L. E., Reynolds, T. L., Cameron, M. A., Conway, M., Lau, E. H. Y., Olsen, J. M., Pavlin, J. A., Shigematsu, M., Streichert, L. C., Suda, K. J., and Corley, C. D. (2015). Using social media for actionable disease surveillance and outbreak management: A systematic literature review. *PLoS One*, 10(10):1–20.

de Vigilância em Saúde, S. (2021). Boletim epidemiológico especial 52: doença pelo coronavírus covid-19. Technical report, Ministério da Saúde.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014.

Gomide, C. S., de Freitas de Carvalho Lima, A. A. T., Gomide, J. S., Roque, D. M., and Silva, T. S. (2014). O twitter como instrumento de detecção de epidemias de dengue e desenvolvimento de políticas públicas. In *XXXVIII Encontro da ANPAD*, pages 1–14.

Jahanbin, K. and Rahmanian, V. (2020). Using twitter and web news mining to predict covid-19 outbreak. *Asian Pacific Journal of Tropical Medicine*, 13(8):378–380.

Kraemer, M. U. G., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D. M., Group, O. C.-. D. W., du Plessis, L., Faria, N. R., Li, R., Hanage, W. P., Brownstein, J. S., Layan, M., Vespignani, A., Tian, H., Dye, C., Pybus, O. G., and Scarpino, S. V. (2020). The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 368(6490):493–497.

Li, J. and Cardie, C. (2013). Early stage influenza detection from twitter. *CoRR*, abs/1309.7340.

MonitoraCovid-19 (2020). O tempo dos dados: explorando a cobertura e oportunidade do sivep-gripe e o e-sus ve. Technical report, Laboratório de Informação em Saúde (LIS), Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT), Fiocruz.

Zhang, Q., Perra, N., Perrotta, D., Tizzoni, M., Paolotti, D., and Vespignani, A. (2017). Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. pages 311–319.