Analisando a associação entre tweets geolocalizados próximos a hospitais e a ocorrência de casos de COVID-19

Bruno Marra de Melo[1], **Fabrício Aguiar Silva**[2]

[1]Instituto de Ciências Exatas e Tecnológicas – Universidade Federal de Viçosa (UFV)
35.690-000 – Florestal – MG – Brasil

Abstract. *The current context of the COVID-19 pandemic has affected the routine of a large part of the population. Major efforts to analyze the pandemic numbers have been raised by researchers around the world. An important challenge is the prediction of the number of confirmed cases in the future, which is useful for decision making of public managers. This work investigates the hypothesis that the Twitter posts with content related to the disease near to hospitals are associated to the number of confirmed cases in the future. To this end, we collected tweets geolocated near hospitals in Brazil during the pandemic, and analyze them through correlation and regression. The results reveal that is possible to estimate the number of confirmed cases in the near future for some capitals, using the geolocated tweets.*

Resumo. *O contexto atual da pandemia de COVID-19 afetou a rotina de grande parte da população. Grandes esforços de análises sobre os números da pandemia foram levantados por pesquisadores de todo o mundo. Um dos desafios enfrentados é previsão de casos confirmados da doença, que pode ser usada para auxiliar nas tomadas de decisões dos gestores públicos. Este trabalho investiga a hipótese de que postagens no Twitter com conteúdo relacionado à COVID-19 e próximos a hospitais estão associados com o número de casos confirmados da doença em algum momento futuro. Para isso, foram coletados tweets geolocalizados próximos a hospitais no Brasil durante a pandemia, e foram feitas análises por meio de correlação e regressão. Os resultados mostraram que, para algumas capitais brasileiras, é possível estimar o número de casos confirmados com base nos tweets.*

1. Introdução

A pandemia de COVID-19 adveio de uma doença que abalou grande parcela da população, e impactou direta ou indiretamente os hábitos, costumes, consumos, dentre vários aspectos da vida humana, que tiveram que ser repensados e reavaliados em vários contextos. A partir dessa situação, esforços em conjunto de milhares de pesquisadores ao redor do mundo foram reunidos, na tentativa de conter, contornar, ou pelo menos frear a contaminação e, consequentemente, perdas ocasionadas pela COVID-19. [Attaallah et al. 2021]

Através da análise de dados, contribuições foram sendo produzidas para mitigar a situação. Grande parte dos trabalhos focaram-se em descobrir padrões no crescimento de contaminações, casos e estimativas para tentar encontrar comportamentos comuns sobre o crescimento de contaminações, casos e mortes ocasionadas pela COVID-19 ao redor do mundo [Muhammad et al. 2020]. O Twitter é uma fonte importante para essas análises.

Por meio dessa rede social, é possível fazer uma análise para extração de dados mais direcionados e relacionados à COVID-19 [Kouzy et al. 2020].

Uma informação potencialmente relevante e até o momento negligenciada nos estudos existentes é a geolocalização das postagens do Twitter. O presente artigo visa investigar se existe uma associação entre tweets relacionados à COVID-19 postados próximos a hospitais com os casos registrados da doença na cidade. A hipótese é que pessoas com suspeitas de estarem contaminadas visitem alguma área hospitalar para exames e consultas, o que levará a uma confirmação (ou não) da contaminação em algum momento no futuro. Para isso, é feito uso da geolocalização em conjunto com dados do Twitter, em um período da pandemia de COVID-19 no Brasil, com dados coletados em regiões próximas a hospitais.

O restante do artigo está organizado da seguinte maneira. A Seção 2 descreve a metodologia e os materiais utilizados. A seção 3 aborda as análises e os resultados obtidos através da pesquisa. Por fim, a seção 4 explicita as conclusões e possíveis evoluções ao presente trabalho.

2. Materiais e Métodos

Para que fosse possível a realização de todas as análises, algumas ferramentas e tecnologias foram essenciais, tanto para a extração, quanto para as fases de manipulação e análise dos dados. No contexto inicial, para que fosse possível a coleta dos dados, o primeiro passo foi a solicitação de uma chave para acesso aos dados do Twitter por meio de uma API REST. Essas podem ser resumidamente definidas por um modelo de arquitetura para sistemas hipermídia, contrastando-os com o restrições de outros estilos de arquitetura, fornecendo uma interface única de conexão. [Arragokula and Ratnam 2016] Uma REST API então pode ser facilmente entendida como uma interface padronizada de transferência de dados, REST (*representational state transfer*) e API (*application programming interfaces*) [Masse 2011].

Para a transferência desses dados por meio do protocolo HTTPS, uma variação do HTTP com mais camadas de segurança, e um baixo impacto em termos de desempenho [Goldberg et al. 1998], foi utilizado o formato JSON (*Javascript Object Notation*). O JSON foi projetado de forma a ser uma linguagem de troca de informações, legível por humanos e fácil para computadores utilizarem. [Schema.org 1999] Para que fosse possível realizar as análises, a linguagem de programação escolhida foi o Python, uma linguagem que pode ser fácil de aprender, possuindo uma sintaxe simples, que pode ser aprendida a partir de conhecimentos básicos, além de ser uma linguagem de alto nível [Python 2001].

Além do Python citado anteriormente, outra ferramenta que foi extremamente importante ao longo do desenvolvimento foi o Jupyter Notebook. Vários projetos e artigos nos últimos tempos foram publicados utilizando o Jupyter para análises e plots de gráficos e resultados [Kluyver et al. 2016]. O Jupyter consiste basicamente em uma ferramenta *open-source*, *browser-based*, que auxilia no desenvolvimento de trechos de código, dados, gráficos e pequenas documentações entre os trechos de código, que podem ser executados individualmente e ou parcialmente, para demonstração de resultados [Randles et al. 2017].

Além das ferramentas supracitadas, algumas bibliotecas foram essenciais para fa-

cilitar a manipulação e ou exibição dos dados do projeto. O Pandas, consiste em uma ferramenta que possibilita a análise de dados de uma forma mais estatística para o Python, uma linguagem de programação de propósitos gerais e científicos [McKinney et al. 2011]. Outra ferramenta de grande importância para realização das análises foi o Matplotlib. O Matplotlib é um pacote gráfico de plotagem e imagem 2D, com finalidade principal em visualização de dados científicos, de engenharia e também financeiros [Barrett et al. 2005].

2.1. Os Dados

O objetivo do trabalho é avaliar a associação entre *tweets* postados próximos a hospitais com os casos de COVID-19 registrados na cidade. Para isso, foram utilizados os seguintes dados:

- **Geometria dos hospitais:** foram coletados do *Open Street Maps* (OSM) os polígonos representando a geometria de 3.190 hospitais de todo o Brasil. Para que essa geometria fosse utilizada para a busca dos *tweets*, cada hospital foi mapeado em uma coordenada central dada pela média aritmética de todo o seu polígono. Isso foi feito para que fosse possível coletar as postagens no Twitter com base em um centro e um raio;
- **Tweets próximos aos hospitais:** com as coordenadas dos hospitais em mãos, foi feita uma coleta geolocalizada usando a API do Twitter durante o período de 09/06/2020 até 21/09/2020. Foram coletados *tweets* nesse período que estivessem geolocalizados dentro de um raio de 500 metros de algum hospital;
- **Casos registrados de COVID-19:** foi utilizada a base de dados disponível em [Álvaro Justen et al 2020] que contempla os casos registrados ao longo do tempo. Foi realizado ainda um processo de geocodificação reversa, para coletar a cidade de cada localização a partir de sua latitude/longitude. Para isso foi usada a API de geocodificação reversa *Nominatim* [Developer 2018].

O passo inicial para a execução do trabalho foi conseguir as chaves necessárias para consulta e coleta dos dados na API do Twitter. Para isso, o Twitter fornece uma API gratuita na v1.1, que permite a extração e busca de *tweets* nos últimos 7 dias. A API permite também 450 requisições a cada 15 minutos, [Twitter 2021] permitindo então que fossem extraídos 100 *tweets*, que é o limite máximo permitido a cada requisição, por localidade nos 7 dias de período. Não se fazia necessária a extração de mais registros visto que a grande maioria ficava abaixo desse número, bem como, a extração já se mostrava demorada devido ao limite de requisições a cada 15 minutos, o que tornava praticamente inviável a extração paginada dos registros, visto que extraindo somente a primeira página, o tempo médio necessário para cada dia de extração era de 2 horas e 30 minutos.

2.2. Preparação dos Dados

Para conseguir um volume de tweets relevantes, foi realizado uma extração exaustiva durante o período de 09/06/2020 até 20/07/2020, em sua v1.0, onde foram armazenados apenas o conteúdo textual dos *tweets*. Nessa versão, eles seguiram um formato que facilitasse sua análise, salvos em um CSV com sua data de coleta, bem como em um formato tabular de latitude, longitude, lista de *tweets*. A Figura 1 exemplifica o formato dessa versão.

Figura 1. Versão 1.0 da coleta dos *tweets*

Antes ainda de trabalhar com os dados coletados, foi implementada uma nova versão de coleta, mantendo todos os dados de retorno do twitter, denotada v2.0, sendo esta então o formato definitivo dos dados. Para essa nova versão, foram coletados dados de 03/08/2020 até 21/09/2020. Nessa nova versão, o formato tabular foi mantido, contudo ao invés de uma lista com o conteúdo do *tweet*, foi armazenado então uma lista de objetos no formato JSON, contendo toda a informação retornada. Não foi possível obter os dados anteriores dado a limitação da chave gratuita fornecida pelo Twitter. Foi criado ainda um arquivo intermediário padronizando as duas versões para início da análise.

Com os dados brutos salvos e armazenados, foi então possível iniciar o seu tratamento. O tratamento dos dados textuais é crucial para o bom entendimento e análise posterior dos mesmos. Existem vários problemas para se fazer o tratamento de dados textuais, em especial tratamento de dados de uma fonte majoritariamente composta por uma linguagem natural bastante informal e não estruturada, como uma rede social. A qualidade de um algoritmo dependerá de quão bem é executada a limpeza de dados. Ao lidar com processamento de linguagem natural, a limpeza de dados fica ainda mais complexa e sensível a falhas [Dukare 2020].

Dessa forma, o conceito de Expressões Regulares (Regex) foi utilizado para remoção de termos e todos os caracteres desnecessários para a interpretação dos dados. Expressão Regular é um padrão usado para descrever uma consulta de pesquisa para extrair informações de um dado conjunto de texto [Dukare 2020]. O primeiro passo de limpeza consistiu na remoção de ruídos. Ruídos podem ser caracterizados como textos que não agregam para a análise. Um exemplo de ruído é por exemplo o link do *tweet*, presente ao final da informação textual de alguns *tweets*. Além dos links, numerais também não eram interessantes para a análise visto que o intuito era correlacionar *tweets* com a COVID-19 e, portanto, os numerais não agregavam informação.

O segundo passo para facilitar a classificação dos *tweets*, foi fazer a remoção das chamadas *stop words*. Palavras irrelevantes adicionam redundância à análise de texto, portanto, ao remover essas palavras, é possível conseguir extrair informações mais apropriadas do texto [Dukare 2020]. Para definição de quais palavras seriam removidas, foi utilizado um banco de palavras[1] de *stop words* em português.

Para determinação de quais *tweets* são relevantes para a análise, foi levantado então um banco de palavras, nos quais a busca foi feita pelo radical da mesma, de forma a classificar um determinado *tweet* como relacionado ao assunto COVID-19. Esse filtro não poderia ser muito rígido, para não reduzir muito a quantidade de dados levantados do twitter, visto que a API foi uma limitação que dificultou bastante a execução do trabalho. O banco de palavras era composto pelas *hashtags* mais populares relacionadas a COVID-19 [Noli da Fonseca et al. 2020], bem como algumas palavras referentes a sintomas, dores, ou correlatas a doença. A Figura 2 mostra com mais detalhes os termos parciais ou completos que foram utilizados para classificação de *tweets* com relevância para análise.

dor	dores	dolorido	sentindo mal	senti mal	covid	corona
febre	tosse	cansado	cansaço	garganta	diarreia	conjutivite
ouvido	esfolação	isolado	trancado	fiqueemcasa	pandemia	isolamentosocial
saude	quarentena	RT-PCR	cotonete	teste	internado	exame
sorologia	imune	paladar	imunocromatografia			

Figura 2. Termos ou palavras classificadoras de *tweets* relevantes

Feito isso, os *tweets* que não eram relevantes foram então eliminados da lista, e um novo arquivo CSV intermediário, somente com os *tweets* relevantes no mesmo formato do arquivo bruto extraído foi gerado. Exemplos de *tweets* classificados foram os mais diversos, como:

'Número mortos novo coronavírus Maranhão sobe ., casos confirmados passam mil... ', 'O mundo prega muitas surpresas. Muitas fazem sofrer, trazem tristeza, dor luto. A partida, tão premat... ', 'Postarei story td dia números atualizados, p/ perguntar: vc feito diminuir dor coleti... ', '#Repost @aliensvshumanos Gripezinha #charge #charges #chargespolíticas #corona #coronavírus #saúde... ', 'Nada dia após outro. Ontem tava sentindo tipo diarreia decorrente ingestão alimentos ven... ', 'Em domingo, senti febre calafrios, jamais imaginei COVID-19, afinal dentro casa direto n... ', 'hoje manhã senti dores fortes abdômen vim consultar exames...'.

Esse então foi insumo para realização das análises subsequentes. Foi preciso converter então o arquivo com esses *tweets*, em um arquivo tabular mais sintetizado, contendo a localização média do hospital (latitude e longitude), bem como o número de *tweets* relevantes para uma data de coleta sumarizada de 7 em 7 dias para fazer uma análise semanal.

Com esses dados sumarizados em mãos, foi possível começar uma análise exploratória sobre os mesmos, para um melhor entendimento dessa distribuição. Um ponto que foi observado já de imediato, foi que a quantidade de *tweets* correlatos em capitais era

[1] Acesso em: https://gist.github.com/alopes/5358189. Acessado: 19/07/2021

mais numerosa que em outras cidades menos populosas, que por consequência, possuíam menos hospitais. A Figura 3 mostra de maneira sumarizada, o número de *tweets* coletados e analisados de todo o Brasil, ao longo de todo o período para elaboração do trabalho.

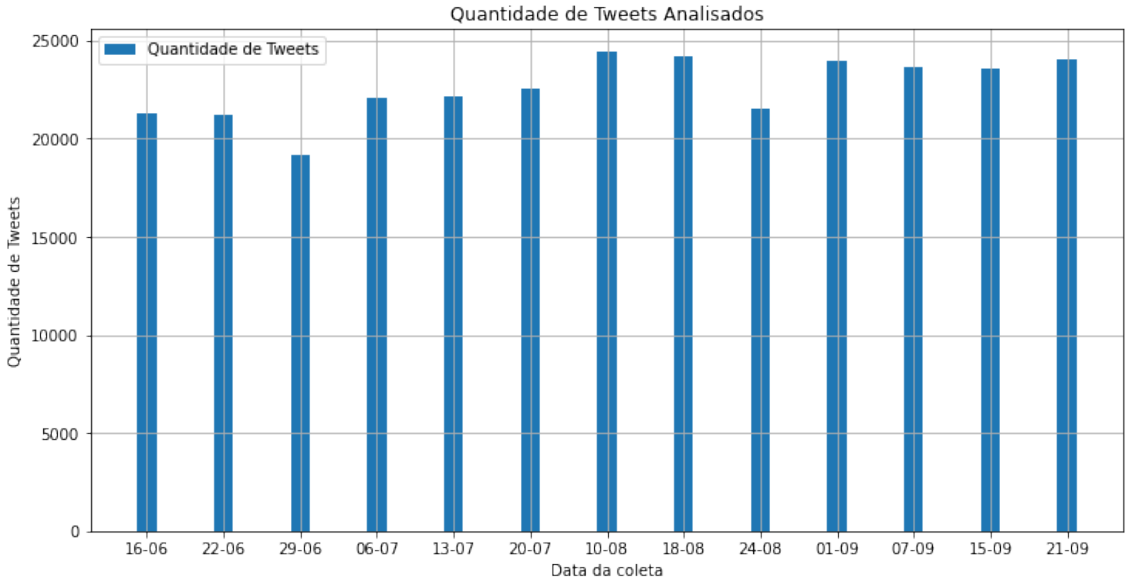

Figura 3. Quantidade de tweets analisados pelos algoritmos

Além da visão do número de tweets analisados, a Figura 4 mostra um gráfico de calor sob todo o território brasileiro, ao longo de todo o período analisado. Nota-se uma clara densidade maior nas regiões sul, sudeste e nordeste em comparação com as regiões centro-oeste e norte. Isso se deve tanto ao número maior de hospitais nessas regiões quanto às características da população das localidades, como poderá ser observado mais adiante no trabalho.

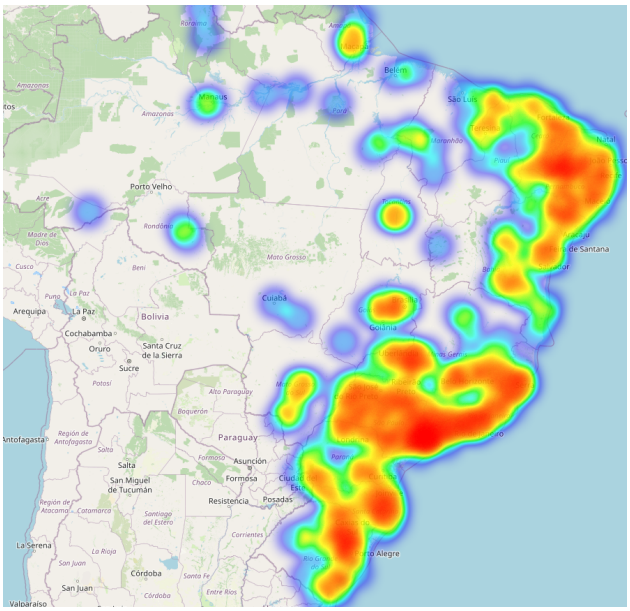

Figura 4. Heatmap de densidade de *tweets* relacionados por região

Para que fosse possível verificar se os dados haviam ou não algum tipo de correlação com os casos de COVID-19 da região, foi necessário coletar dados confiáveis de casos por localidade [Álvaro Justen et al 2020].

3. Resultados e Análises

O processo para obtenção das melhores correlações entre os casos confirmados de COVID-19 em uma cidade e o número de *tweets* geolocalizados próximos a hospitais dessa cidade passou por muitas análises para se alcançar os resultados. Inicialmente, foi necessário selecionar cidades com dados suficientes para refletir melhor a correlação entre os casos. Foram selecionadas então cidades em pontos distintos do país, com realidades diferentes para o estudo. Dessa forma, optou-se por capitais, nas quais o número de hospitais eram maiores, possuem maior área de cobertura bem como população. Foram selecionadas as cidades de *São Paulo, Belo Horizonte, Rio de Janeiro, Fortaleza, Porto Alegre, Recife, Manaus e Salvador*.

3.1. Correlações

Inicialmente, foi realizada uma análise da correlação entre os casos confirmados de COVID-19 e a quantidade de *tweets* relacionados à doença. Cada variável aleatória (X_i) na tabela de correlação é correlacionada com cada um dos outros valores na tabela (X_j). Isso permite que seja possível identificar quais pares têm a correlação mais alta entre si [Glen 2016].

Dentre as 8 cidades selecionadas, 3 tiveram uma correlação acima de 60% entre os *tweets* da semana atual, refletindo nos casos de COVID-19 em alguma semana seguinte. Para que fosse possível chegar a esses resultados, foram analisadas as semanas durante todo o período, variando de 30/06/2020 até 21/09/2020.

Para que os dados não ficassem discrepantes, foi necessário normalizar utilizando *Min/Max* tanto sobre os casos de COVID-19 quanto sobre os *tweets* relacionados a doença na cidade analisada, para que os resultados fossem coerentes. Ao fazer isso, todos os números são transformados no intervalo $[0, 1]$, o que significa que os valores mínimo e máximo de uma variável serão 0 e 1, respectivamente [Loukas 2020]. A equação que representa essa normalização pode ser expressa da seguinte maneira:

$$x_{scaled} = \frac{x - min(x)}{max(x) - min(x)} \tag{1}$$

As Tabelas das Figuras 5 e 6 exibem, respectivamente, os resultados obtidos das correlações para as cidades analisadas, testando todas as combinações existentes e analisando, tanto os casos de COVID-19 corresponderem aos *tweets* de $n \in \{1, 2, 3, 4\}$ semanas seguintes, quanto o oposto, ou seja, os *tweets* da semana analisada corresponderem aos casos da doença $n \in \{1, 2, 3, 4\}$ semanas seguintes. O valor em vermelho reflete a melhor correlação obtida para a cidade variando a semana avaliada. As duas análises foram realizadas buscando verificar qual traria um resultado mais adequado; contudo, a hipótese mais clara e que o presente artigo buscou analisar é a segunda, onde os *tweets* da semana corrente correspondem aos casos de COVID-19 nas semanas seguintes. Ou seja, o usuário realiza uma publicação próximo a um hospital, talvez por estar com sintomas e

indo realizar um teste, mas o teste é confirmado apenas algum tempo depois. Vale ressaltar que, para casos onde houve uma alta correlação negativa, ela foi considerada a melhor, visto que uma alta correlação negativa representa inversamente a proporção de casos com *tweets*, sendo mais interessante para a análise.

	mesma semana	1 semana depois	2 semanas depois	3 semanas depois	4 semanas depois
São Paulo	0,24	0,38	0,26	0,75	0,25
Rio de Janeiro	0,44	0,26	0,19	0,17	0,33
Belo Horizonte	0,62	0,65	0,56	0,38	-0,10
Fortaleza	0,36	0,40	0,49	0,34	0,29
Porto Alegre	-0,07	0,11	-0,02	-0,34	-0,61
Recife	-0,05	0,21	-0,09	-0,52	0,12
Manaus	0,27	0,25	0,31	0,36	0,50
Salvador	-0,05	0,00	0,13	0,09	0,24

Figura 5. Correlações entre casos de COVID-19 x *tweets* - semanas

	mesma semana	1 semana depois	2 semanas depois	3 semanas depois	4 semanas depois
São Paulo	0,24	0,21	0,55	0,01	0,69
Rio de Janeiro	0,44	0,29	0,03	-0,38	-0,60
Belo Horizonte	0,62	0,61	0,34	0,30	-0,33
Fortaleza	0,36	0,23	0,63	0,30	-0,15
Porto Alegre	-0,07	-0,26	-0,17	-0,62	-0,32
Recife	-0,05	0,31	0,45	0,58	0,22
Manaus	0,27	0,14	-0,04	0,40	0,41
Salvador	-0,05	0,21	0,33	-0,05	0,12

Figura 6. Correlações entre *tweets* x casos de COVID-19 - semanas

Os gráficos da Figura 7 explicitam com detalhes essa comparação feita para as capitais analisadas. Para cada cidade, é apresentado o melhor cenário (qual a semana n no futuro que melhor se correlaciona com os *tweets* atuais), visto que os resultados individuais para cada cidade foram diferentes. Vale ressaltar que, como as melhores correlações obtidas para cada capital foram comparando os *tweets* da semana com casos semanas diferentes, para cada semana a mais ocorre a perda de uma semana de análise. Destaque para as cidades de São Paulo, Belo Horizonte e Fortaleza, que possuem correlações acima de 60% positivas para o período, demonstrando um comportamento previsível para o número de casos de COVID-19.

A tabela 8 apresenta algumas características de cada cidade, sendo: o número de hospitais analisados, o IDHM[2] do município, sua população, área e densidade demográfica.

Com isso, o número de hospitais analisados pode explicar a baixa correlação para as cidades com poucos hospitais, reduzindo a área de busca e análise por exemplo, bem como demonstrado no caso de Manaus, em que quase metade do período sem incidência de tweets relacionados, contando com a pior correlação em conjunto com a cidade de Salvador. Outro ponto ainda bastante relevante sobre Manaus, é que essa cidade tem o

[2]Índice de Desenvolvimento Humano Municipal

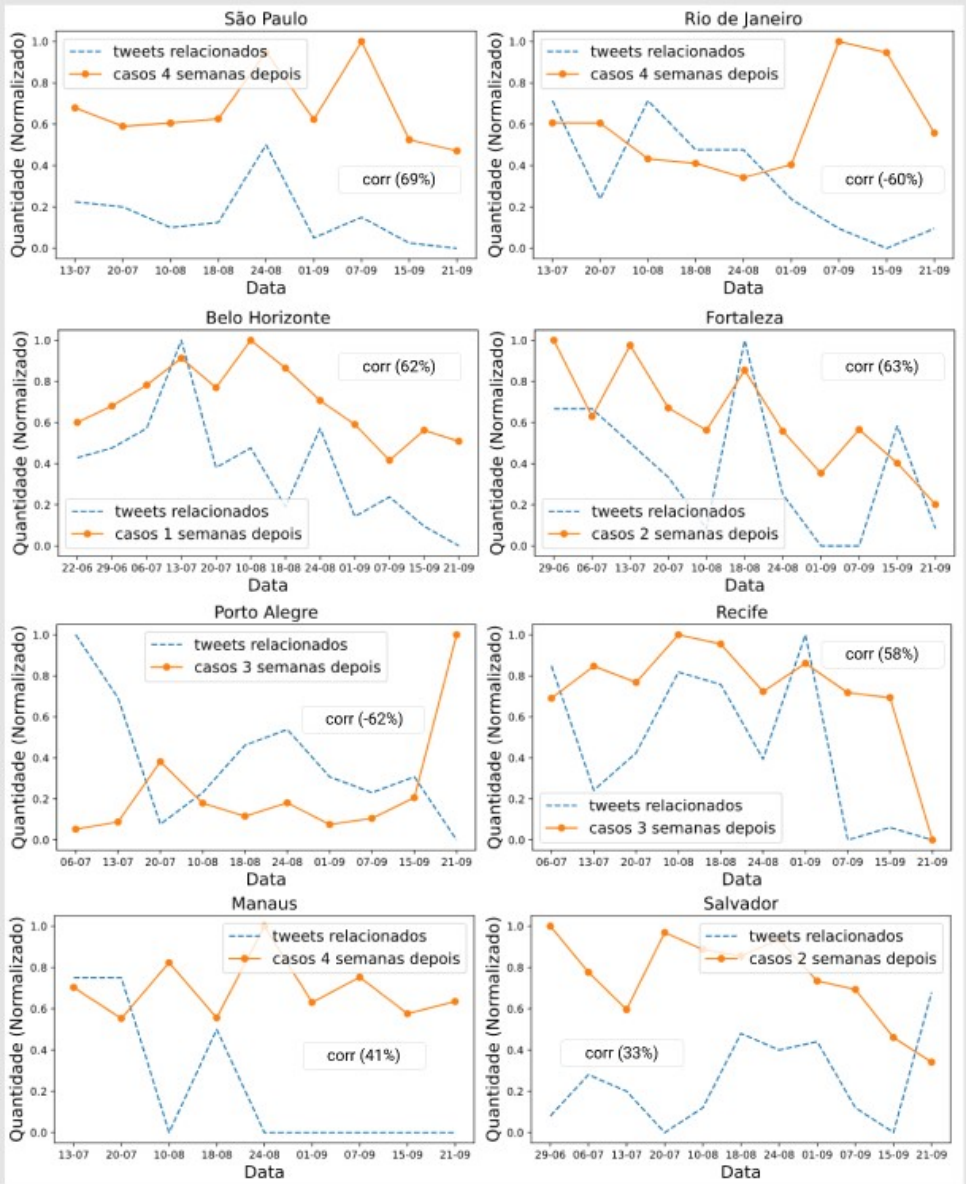

Figura 7. Melhores correlações para as capitais analisadas

menor número de hospitais analisados e de longe a maior extensão territorial em Área das capitais, reduzindo então a precisão dos dados como foi possível observar.

Outro ponto importante a ser observado, a partir dos dados coletados pelo IBGE para o ano de 2010, cidades com o IDHM em conjunto com uma densidade populacional maior, tiveram tendência ao comportamento de *tweets* relacionados representarem melhor os casos de COVID-19.

3.2. Modelos Regressivos

Após o estudo e análise feitos sobre as correlações supracitadas, foram elaborados testes utilizando técnicas de regressões, de forma a obter um modelo preditivo que conseguisse reproduzir corretamente a previsibilidade dos casos de COVID-19 de acordo com a data e os respectivos *tweets* relacionados a doença. Em todas as análises, foi gerado um modelo

	N. de Hospitais	IDHM	Densidade Demográfica (hab/km²)	População	Área (km²)
São Paulo	100	0,805	7398,26	12.396.372	1521,110
Rio de Janeiro	106	0,799	5265,82	6.775.561	1200,329
Belo Horizonte	42	0,810	7167,00	2.530.701	331,354
Fortaleza	35	0,754	7786,44	2.703.391	312,353
Porto Alegre	27	0,805	2837,53	1.492.530	495,390
Recife	60	0,772	7039,64	1.661.017	218,843
Manaus	18	0,737	158,06	2.255.903	11401,092
Salvador	27	0,759	3859,44	2.900.319	693,453

Figura 8. Informações demográficas dos municípios

individual para cada uma das 8 capitais analisadas, considerando a semana na qual se obteve a melhor correlação para a cidade. Foram testadas 3 técnicas: regressão linear, regressão logística e algoritmo genético.

A Regressão Linear consiste na tentativa de encontrar uma função representando uma reta, dado um conjunto de pontos de dispersão. O objetivo é minimizar a distância entre os pontos e a reta, buscando representar a melhor função que descreve a sequência dos dados dispersos ao longo do tempo, de forma a tentar predizer o comportamento de alguma variável independente [Maroco 2003].

Essa tentativa se mostrou pouco eficiente para representação de todas as capitais. O maior *score* (R^2) obtido foi para a cidade de Fortaleza, com 0,73. O score de um modelo varia entre 0 e 1, onde 0 indica que a variação dos casos não é explicada pelas publicações no Twitter, e 1 que 100% da variação dos casos é explicada pelos *tweets*. Em outras palavras, indica que a reta passa exatamente por todos os pontos.

Após o modelo de Regressão Linear, foi analisado também a possibilidade de aplicação de uma Regressão Logística ao problema, considerando as melhores correlações para as capitais analisadas. O modelo de Regressão Logística explora o uso de classes, dessa forma a função deixa de retornar um valor exato e retorna em qual classe o dado analisado se enquadra. a Regressão Logística também se diferencia da Regressão Linear, em que o método dos mínimos quadrados não é interessante para resolução do problema. Sendo assim, os valores que a variável dependente assume pode possuir valor nominal ou ordinal [Figueira 2006]. Considerando que o problema se trata de variáveis binominais (semana do ano + *tweets* na semana) X (casos na semana), a Regressão Logística Nominal permitiu que os resultados melhorassem de forma significativa. Para essa técnica, todas as cidades obtiveram um *score* acima de 0,60 com algumas cidades alcançando valores próximos a 0,80.

Por fim, foi testada também uma terceira estratégia de predição composta por uma combinação entre um algoritmo genético com uma função exponencial euleriana, ajustando seus parâmetros para obter a melhor curva possível para a distribuição analisada. A equação base que passou a sofrer modificações pode ser expressa da seguinte maneira:

$$y = \frac{1,0}{1,0 + e^{(-a(xb))}} + deslocamento \tag{2}$$

Nessa função, os parâmetros a e b são ajustados por meio do algoritmo genético, criando gerações aleatórias para tentar obter a melhor combinação de a e b que refletem a curva dos dados dispersos. A implementação utilizada foi a *Differential Evolution* da biblioteca *scipy* do Python. Essa estratégia garante uma pesquisa completa sobre o espaço dos parâmetros, de acordo com os limites de pesquisa. No código desenvolvido, esse limite é definido pelos valores máximos e mínimos dos dados dispersos [Phillips 2018]. Essa escolha foi tomada baseando na facilidade de implementação da mesma, além dos parâmetros da função auxiliarem nos resultados obtidos.

Para revisão e análise de todos os algoritmos lado a lado, a Figura 9 mostra os resultados de cada estratégia para cada uma das cidades analisadas, baseando-se na métrica R^2, visto que é uma métrica avaliativa que pode representar a qualidade de um modelo de regressão, ou seja, a diferença entre o ponto da curva gerada e os pontos dispersos.

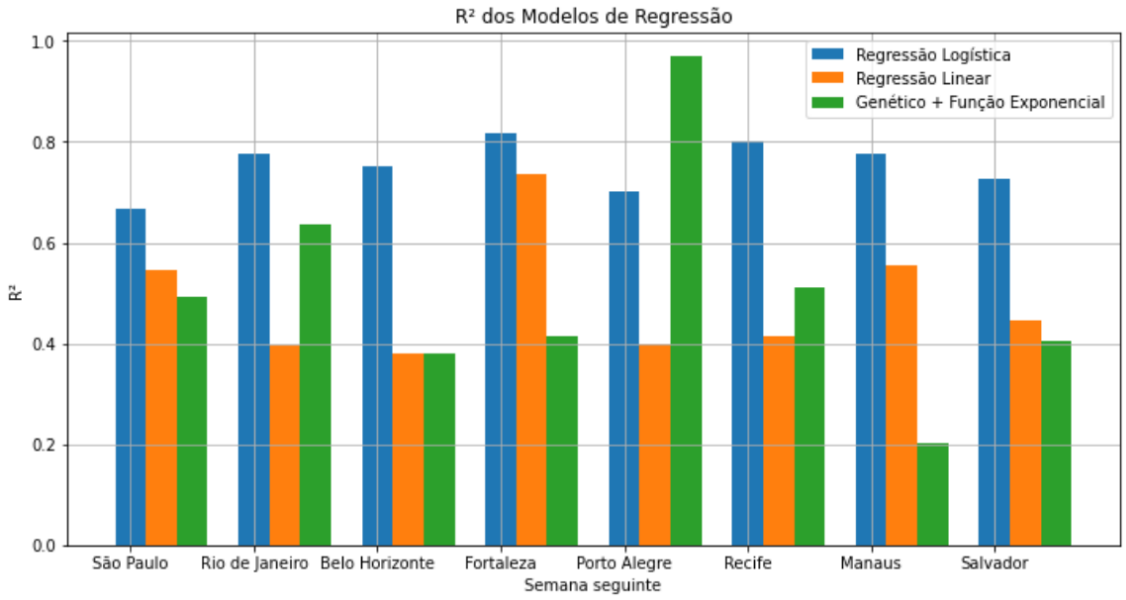

Figura 9. Scores gerados para os modelos discutidos

Pode-se observar que o modelo de Regressão Logística se mostrou muito consistente na representação da grande maioria das cidades, sendo inclusive possivelmente viável para uma implementação única desconsiderando a cidade a ser analisada. Apesar disso, o modelo utilizando a função exponencial euleriana representou quase que exatamente o comportamento para o período analisado na cidade de Porto Alegre, podendo ser bastante significativo para um modelo preditivo para essa cidade especificamente, visto que seu comportamento se assemelhou muito com uma função exponencial $1/c^x$ onde c é uma constante. O fato de ter possuído uma maior correlação negativa, em conjunto com Rio de Janeiro que também se assemelha nesse aspecto, compactua com os dois melhores desempenhos para o modelo de função exponencial combinada ao algoritmo genético; inclusive as funções geradas por ambos possuem comportamentos relativamente similares.

Além do R^2, também foi calculado o RMSE (Raiz quadrada do erro-médio) para cada um dos modelos. O RMSE é a medida que calcula a raiz quadrática média dos erros entre valores reais e as predições obtidas [Rezende 2018]. Nesse caso, quanto menor

for o RMSE, maior previsibilidade possui o modelo. A Figura 10 exemplifica com mais detalhes os valores obtidos assim como feito para o R^2.

Pode-se perceber então, que para o caso da Regressão Logística, o RMSE se destacou de forma negativa mediante aos demais modelos. Isso pode ser explicado pela característica do modelo de regressão logística, que normalmente é bastante eficiente para modelos de classificação, onde existe um número limitado de classes e cada ponto é classificado em uma delas. Por exemplo, em um modelo regressivo onde se analisa peso e altura, o indivíduo é classificado como obeso ou não. No caso desse modelo, foi utilizado então a função *LabelEncoder*, do pacote de pré-processamento da biblioteca *sklearn*, para separar o eixo Y em classes, correspondente aos casos reais de COVID-19, sendo o eixo X uma composição de semana e *tweets*. Esta, cria um número de classes entre 0 e quantidade de dados em Y - 1, possuindo suporte para fazer o *fit* e ainda retornar os valores codificados. Na predição, ao ser informado uma semana e um número de *tweets* relacionados, o modelo retorna em qual das classes ele mais se aproxima. Cada classe possuindo um range de casos de COVID-19. Portanto, o RMSE é maior, visto que como explicado calcula a raiz quadrática média dos erros [Rezende 2018], e da mesma forma se explica um R^2 maior, visto que mais indivíduos vão ser classificados corretamente pela métrica R^2 no modelo. Vale ressaltar que a métrica não diz respeito necessariamente a qualidade do modelo em si, mas o quão longe do valor exato o valor obtido pelo modelo se encontra, e, portanto, não inviabiliza o uso da Regressão Logística para a solução do problema.

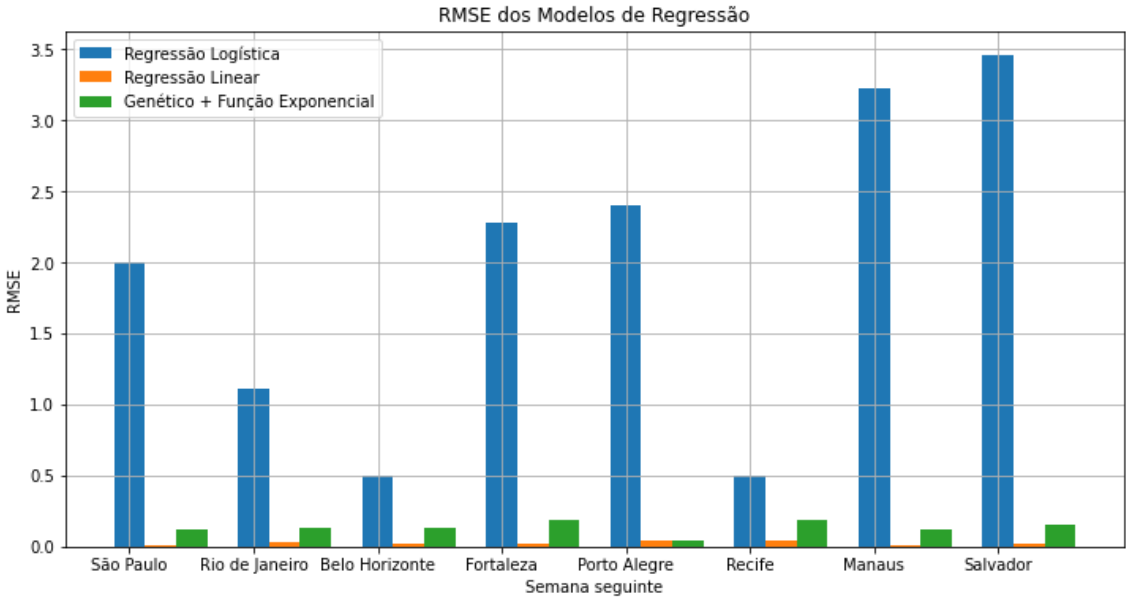

Figura 10. RMSE's gerados para os modelos discutidos

3.3. Análise

Com os resultados, foi possível observar que os *tweets* geolocalizados próximos a hospitais podem ser uma fonte significativa de informação para a previsão de casos de COVID-19, e representar de forma coerente o comportamento de uma região, mesmo que nem todo mundo faça uso do Twitter para expressar sentimentos. Os resultados refletem de forma consistente uma micro-região, como foi o caso desse trabalho, buscando as micro-regiões

próximas a hospitais para representação de casos de COVID-19, se mostrando especialmente eficiente em algumas capitais, que são cidades mais populosas e com maior número de hospitais. Apesar disso, um modelo genérico para todo o país não será muito eficiente, dadas as características individuais de cada região, estado e ou cidade, sendo interessante uma análise em regiões menores (como cidade, por exemplo) para obtenção de resultados representativos.

Com os resultados deste trabalho, é possível criar modelos geolocalizados que, com base nas postagens no Twitter, serão capazes de estimar os casos de alguma doença contagiosa. Isso poderá auxiliar as tomadas de decisões, e o planejamento de recursos, em futuras pandemias.

4. Conclusões e Trabalhos Futuros

Como próximos passos, pode ser expandido o período de coleta dos dados, para enriquecer as análises e verificar o comportamento do modelo ao longo de todo o período pandêmico. Outra possibilidade seria realizar as análises para outras cidades, tanto de pequeno quanto grande porte fazendo um estudo comparativo entre elas, verificando se existem grandes diferenças entre as informações obtidas neste artigo.

Por fim, algumas possibilidades de projetos pilotos podem ser implementados, como por exemplo na cidade de Porto Alegre, onde a curva obtida pelo modelo de função exponencial euleriana foi bastante similar ao comportamento dos casos. Além deste, um projeto piloto genérico utilizando algum dos modelos e estratégias discutidas nesse artigo pode se mostrar bastante eficiente se possuir o mesmo comportamento para outras cidades, podendo ser bastante útil na tentativa de mitigar impactos pandêmicos futuros.

Referências

Arragokula, S. and Ratnam, M. Y. (2016). Architectural styles and the design of network-based software architectures. *International Journal of Ethics in Engineering & Management Education*.

Attaallah, A., Ahmad, M., Seh, A. H., Agrawal, A., Kumar, R., and Khan, R. A. (2021). Estimating the impact of covid-19 pandemic on the research community in the kingdom of saudi arabia. *Computer Modeling in Engineering & Sciences*, 126(1):419–436.

Barrett, P., Hunter, J., Miller, J. T., Hsu, J.-C., and Greenfield, P. (2005). matplotlib–a portable python plotting package. In *Astronomical data analysis software and systems XIV*, volume 347, page 91.

Developer, M. (2018). Open search (nominatim) api. `https://developer.mapquest.com/documentation/open/nominatim-search/#:~:text=Reverse%20Geocode&text=This%20is%20the%20process%20where,or%20with%20the%20OpenStreetMap%20ID`. Acessado em: 14/07/2021.

Dukare, A. K. (2020). Data cleaning for nlp of social media data in 2 simple steps. `https://towardsdatascience.com/data-cleaning-for-nlp-of-social-media-text-in-2-simple-steps-6ca48fa99c17`. Acessado em: 12/07/2021.

Figueira, C. V. (2006). Modelos de regressão logística. In *Programa de Pós-Graduação em Matemática*, page 68. Universidade Federal do Rio Grande do Sul.

Glen, S. (2016). Correlation matrix: Definition. https://www.statisticshowto.com/correlation-matrix/. Acessado em: 18/07/2021.

Goldberg, A., Buff, R., and Schmitt, A. (1998). A comparison of http and https performance. *Computer Measurement Group, CMG98*, 8.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J. B., Grout, J., Corlay, S., et al. (2016). *Jupyter Notebooks-a publishing format for reproducible computational workflows.*, volume 2016. 20th International Conference on Electronic Publishing.

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. (2020). Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter. *Cureus*, 12(3).

Loukas, S. (2020). Everything you need to know about min-max normalization: A python tutorial. https://towardsdatascience.com/everything-you-need-to-know-about-min-max-normalization-in-python-b79592732b79. Acessado em: 18/07/2021.

Maroco, J. (2003). *Análise Estatística – Com utilização do SPSS*. 2ª edição; Edições Sílabo.

Masse, M. (2011). *REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces*. "O'Reilly Media, Inc.".

McKinney, W. et al. (2011). pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, 14(9):1–9.

Muhammad, L., Islam, M. M., Usman, S. S., and Ayon, S. I. (2020). Predictive data mining models for novel coronavirus (covid-19) infected patients' recovery. *SN Computer Science*, 1(4):1–7.

Noli da Fonseca, M., Santos Accioly, N., Garcias, C., and Ferentz, L. (2020). Hashtags relacionadas à covid-19 no brasil: utilização durante o início do isolamento social — hashtags related to covid-19 in brazil: the usage during the beginning of the social isolation. *Com. Ciências Saúde 2020;31 Suppl 1:131-143*, page 135.

Phillips, J. (2018). Regressão não linear com python - qual é um método simples para ajustar melhor esses dados? https://www.ti-enxame.com/pt/python/regressao-nao-linear-com-python-qual-e-um-metodo-simples-para-ajustar-melhor-esses-dados/805615633/. Acessado em: 08/08/2021.

Python, I. (2001). *Python*. https://www.python.org.

Randles, B. M., Pasquetto, I. V., Golshan, M. S., and Borgman, C. L. (2017). Using the jupyter notebook as a tool for open science: An empirical study. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–2. IEEE.

Rezende, T. (2018). Rmse ou mae? como avaliar meu modelo de machine learning? https://pt.linkedin.com/pulse/rmse-ou-mae-como-

avaliar-meu-modelo-de-machine-learning-rezende. Acessado em: 07/09/2021.

Schema.org (1999). Introducing json. https://www.json.org/json-en.html. Acessado em: 31/05/2021.

Twitter (2021). Search tweets: Standard v1.1. https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets. Acessado em: 05/06/2021.

Álvaro Justen et al (2020). Brasil.io covid-19. https://brasil.io/dataset/covid19/boletim/. Acessado em: 14/07/2021.