Covid-19 e Tweets no Brasil: coleta, tratamento e análise de textos com evidências de estados afetivos alterados em momentos impactantes

Mario Maia[1], **Erneson Oliveira**[1], **Luciano Gallegos**[1]

[1]Universidade de Fortaleza – Unifor
60.811-905 – Fortaleza – CE – Brasil

mariogmp@edu.unifor.br, luciano.gallegos@unifor.br

Abstract. *The Twitter receives many texts (tweets) daily shared during relevant events such as natural disasters, accidents, and even disease outbreaks, allowing researchers from different areas to study and predict relationships between people, places, and the behavior of diseases spreading in cities, states, and countries. These tweets, once related to the New Coronavirus (Covid-19) pandemic in Brazil, could bring insights at relevant moments of this event. In order to identify and analyze these moments, we collect, clean and select tweets over 90 days in the first months of 2020 to explore trends of sentiments shared by people during the Covid-19 pandemic within Brazil. In these tweets, we found evidence of affective state changes along relevant moments, such as the first confirmed case and the first death by the Covid-19 in Brazil, observing high negative and positive scores before and the neutral stabilization of these sentiments after each impacting moment.*

Resumo. *O Twitter recebe diariamente muitos textos (tweets) compartilhados durante eventos relevantes como desastres naturais, ocorrências de acidentes, e surtos de doenças, permitindo a pesquisadores de diversas áreas estudar e prever as relações existentes entre pessoas, locais, e até mesmo as características do espalhamento de doenças em cidades, estados, e países. Estes tweets, quando relacionados à pandemia do Novo Coronavírus (Covid-19) no Brasil, podem trazer intuições de tendência e comportamentos em momentos relevantes deste evento. Visando identificar e analisar estes momentos, coletamos, tratamos e selecionamos tweets ao longo de 90 dias, nos primeiros meses do ano de 2020, para explorar tendências de sentimentos compartilhadas por pessoas, durante a epidemia de Covid-19 no Brasil. Nestes tweets, encontramos evidências de mudança dos estados afetivos analisados em momentos relevantes, como o primeiro caso confirmado e a primeira morte no Brasil pelo Covid-19, observando altas negativas e positivas antes e a estabilização desses sentimentos após momentos de grande impacto.*

1. Introdução

O Twitter, uma das plataformas de mídia social mais famosas do mundo, recebe diariamente muitos textos durante eventos relevantes como desastres naturais, ocorrências de acidentes, e até mesmo surtos de doenças. Na prática, pesquisadores de diversas áreas, como das ciências sociais e da epidemiologia, vêm associando estes textos para estudar e

prever possíveis relações existentes entre pessoas, locais e características do espalhamento de doenças entre regiões [Lopreite et al. 2021, Jiang et al. 2020].

Recentemente, alguns trabalhos com foco de tweets vêm sendo relacionados ao Novo Coronavírus (Covid-19), uma doença que originou-se no final de 2019 na cidade de Wuhan, na China, e foi declarado uma emergência de saúde pública (pandemia) de interesse internacional em 30 de janeiro de 2020 pela Organização Mundial da Saúde [Xiang et al. 2020]. Em um deles, é apresentado uma metodologia para captar os principais tópicos em discussão no Twitter para analisar o impacto da pandemia COVID-19 no Brasil durante os primeiros meses de 2020 para analisar os principais temas e sentimentos relacionados durante o período [de Melo and Figueiredo 2021]. Em outro, é realizado uma análise do conteúdo de tweets coletados entre 19 de março e 3 de abril de 2020 de tópicos do sentimento do usuário, informando o predomínio de sentimentos negativos, como medo, tristeza e raiva [Rodrigues de Andrade et al. 2021]. Apesar de relevantes, nenhum destes trabalhos explorou tendências em períodos anteriores e posteriores em marcos relevantes da epidemia de Covid-19 no Brasil, como o primeiro caso confirmado no final de fevereiro de 2020 [1] e a primeira morte, ocorrida em meados de março de 2020.

Neste artigo, coletamos tweets ao longo de 90 dias nos primeiros meses de 2020, objetivando explorar tendências positivas, negativas, e neutras de sentimentos compartilhadas por pessoas, em períodos anteriores e posteriores de marcos relevantes da epidemia de Covid-19 no Brasil, tais como o primeiro caso confirmado e a primeira morte. O artigo está organizado da seguinte forma: na Seção 2, descrevemos as etapas para captar, tratar e selecionar tweets cuja localização esteja no território brasileiro; na Seção 3 fazemos a comparação entre dois analisadores de sentimentos e os critérios utilizados para a escolha de um deles, e na Seção 4 apresentamos análises exploratórias e temporais dos tweets selecionados, levando em consideração pontuações (scores) de análise de sentimento obtidas anteriormente. Por último, encerramos este artigo com nossos comentários finais na Seção 5.

2. Metodologia

Os textos do Twitter utilizados neste trabalho foram obtidos por meio da base de dados GeoCov19 [Qazi et al. 2020]. Esta base contém mais de 524 milhões de tweets com hashtags e palavras-chave relacionados à pandemia de Covid-19, publicados durante 91 dias e com início no dia 1º de fevereiro de 2020. O download dos tweets foi feito manualmente no site do GeoCov19. Estes tweets estão escritos em diversos idiomas e armazenados de forma compacta em formato ZIP. Os arquivos obtidos foram salvos em mídia física SSD, totalizando aproximadamente 270 GB de espaço ocupado, particionados em 91 arquivos, correspondentes aos 91 dias da coleta dos dados.

Os tweets existentes na base de dados do GeoCov19 possuem localização, o que pode ser informado pelo próprio usuário em seu perfil, por localizações sugeridas pelo Twitter no ato da submissão do tweet, por meio de coordenadas de GPS fornecidas pelo aparelho computacional do usuário, ou por meio de topônimos existentes no corpo do texto do tweet. Estes tweets são disponibilizados em formato JSON e possuem os seguinte campos: *tweet_id* (identificador do tweet), *created_at* (data e hora da

[1] `https://coronavirus.saude.gov.br/linha-do-tempo/`

criação do tweet), *user_id* (identificador do usuário), *geo_source* (informa qual campo de localização possui a informação mais precisa), *geo* (localização do aparelho computacional do usuário), *place* (localização sugerida pelo Twitter), *user_location* (localização informada pelo usuário, em seu perfil) e *tweet_locations* (uma ou mais localizações provenientes de topônimos do texto do tweet). Os campos de localização possuem os seguintes atributos: *country_code* (país), *state* (estado), *county* (condado) e *city* (cidade).

Na base de dados do GeoCov19, o campo *geo_source* indica quatro valores possíveis sobre tweets, de acordo com a sua respectiva precisão em relação a sua real localização, aqui listados de forma decrescente: *coordinates*, *place*, *user_location* e *tweet_text* que referenciam, respectivamente, os campos *geo*, *place*, *user_location* e *tweet_locations*. O campo *geo* contém a localização mais precisa [Qazi et al. 2020] e o campo *tweet_locations*, a menos precisa, pois pode apresentar uma ou mais localizações, não permitindo a confirmação da real localização do tweet.

A descompactação, extração e armazenamento dos tweets relativos a estados e cidades brasileiras foram obtidos por meio de algoritmo de execução automática, por nós desenvolvido. Os arquivos de cada dia em formato compactado ZIP, provenientes da base de dados do GeoCov19, são descompactados, resultando em 91 arquivos JSON. Ao total, os tweets e seus campos descompactados ocupam aproximadamente 1,2 GB de espaço em disco rígido.

A cada tweet descompactado, os registros nele contidos são validados quanto à localização, a fim de se verificar se pertencem a uma localização brasileira por meio do campo de localização mais preciso (indicado pelo atributo geo_source). A partir desta localização, verificamos se o atributo *country_code* refere-se ao Brasil e se os valores *city* e *state* (cidade e estado) foram informados. O campo *county* (condado) não foi considerado, por se tratar de uma divisão territorial não utilizada no país. Caso a opção do campo de localização considerado for o *tweet_locations*, que pode possuir mais de uma localização, é verificado se ao menos uma delas é brasileira. Os tweets descompactados e extraídos, cujos campos de localização sejam brasileiros, são armazenados em novos arquivos JSON, totalizando 91 arquivos com 6.288.254 tweets ao total. Para cada arquivo JSON, são armazenados também arquivos CSV contendo seus respectivos identificadores de tweets, para serem utilizados no processo de hidratação (*hydrate*) que será abordado adiante.

Devido ao grande volume de dados obtidos, armazenamos os registros brasileiros em formato JSON utilizando o banco de dados MongoDB, em uma instância local de banco de dados, contendo índices nos atributos referentes aos registros armazenados, visando proporcionar maior agilidade nas consultas. O MongoDB é um software de banco de dados orientado a documentos, de código aberto e multiplataforma, classificado como um software NoSQL e que permite a utilização de documentos semelhantes a JSON com esquemas [Győrödi et al. 2015]

O método de *hydrate* é útil para obter os detalhes de uma coleção de tweets, permitindo a obtenção dos textos de cada tweet extraído e que não estão originalmente na base de dados do GeoCov19. Utilizando estratégia similar a [Chen et al. 2020], aplicamos a ferramenta Twarc para acessar a API do Twitter e efetuar as buscas de tweets em sua base de dados. No Twarc, é necessário a configuração de credenciais de acesso à API,

que podem ser solicitadas na página de desenvolvedores do Twitter [2] e será analisada pela equipe do Twitter baseada nos objetivos da utilização da API. Uma vez obtida as credenciais, elas devem ser configuradas no Twarc pelo comando **twarc configure**, permitindo a execução do *hydrate* para todos os arquivos CSV contendo identificadores de tweets gerados anteriormente.

As execuções dos *hydrates* foram realizadas manualmente via linha de comando. Os arquivos JSON resultantes totalizaram aproximadamente 33,2 GB de espaço em disco. Uma vez finalizadas estas execuções, é computada automaticamente a leitura dos arquivos JSON para obtenção dos textos dos tweets e os respectivos idiomas em que foram escritos, obtidos através dos atributos *full_text* e *lang* providos pelo Twitter.

Durante a realização da leitura dos arquivos JSON, os tweets contidos no banco de dados MongoDB são atualizados com estes novos dados obtidos, incluindo então os atributos *text* e *lang* e perfazendo um total de 5.104.973 registros. Aqueles registros que não tiveram o texto ou o idioma retornados pelo *hydrate* foram desconsiderados. Desta forma, o banco de dados conterá os seguintes atributos: *tweet_id*, *created_at*, *user_id* e *geo_source*, provenientes do GeoCov19, os atributos *state* e *city* decorrentes do processamento realizado nos demais atributos do GeoCov19 referentes às localizações encontradas para cada tweet, e os atributos *text* e *lang* após a aplicação do método *hydrate*.

2.1. Exploração e seleção de dados

Após o hydrate ser utilizado para a obtenção de textos dos tweets, realizamos uma exploração nos dados obtidos a fim de se investigar os resultados em relação à confiabilidade de suas localizações, ou seja, queremos verificar se a localização registrada do tweet pode ser considerada confiável para as nossas análises. Desta forma, investigamos os tweets pelo atributo *geo_source* e verificamos que mais de 60% dos tweets tiveram *tweet_text* como origem de localização. Neste caso, a localização destes tweets foram extraídas a partir de topônimos encontrados no corpo do texto e, portanto, são menos confiáveis. Investigamos também os registros referentes a *tweet_text* e, de fato, encontramos inconsistências entre a localização definida para o tweet e as palavras extraídas. As inconsistências encontradas consistiam em localizações extraídas de substantivos comuns, não-topônimos, e localizações extraídas de topônimos mas que não correspondiam à localização real, quando comparadas ao contexto da frase. Por este motivo, não consideramos os registros de tweets cuja origem seja baseada em *tweet_text* nas análises realizadas posteriores.

Ao explorar os registros por idioma, notamos que os registros em português totalizavam apenas 23% dos tweets. Comparado a tweets escritos no idioma espanhol, notamos que a cidade de São Paulo possuíam o maior número de registros deste tipo, e que somente 4% dos registros desta cidade indicavam o português como idioma. A grande quantidade de registros não escritos em idioma português encontrados na cidade de São Paulo levou-nos a desconsiderar tweets em língua estrangeira, e a focar em tweets em língua portuguesa. Observamos também que no primeiro dia de coleta realizada pelo GeoCov19 (1^o de fevereiro de 2020) todos os registros com regiões do Brasil indicavam o idioma inglês. Os tweets em língua portuguesa só passaram a ocorrer a partir do segundo dia, fazendo com que o conjunto de registros brasileiros compreendesse, então, o

[2] `https://developer.twitter.com/`

período de 2 de fevereiro de 2020 a 1º de maio de 2020. Assim, temos 90 dias a partir de 2 de fevereiro de 2020 e apenas os tweets no idioma português, em um total de 1.219.482 registros selecionados.

2.2. Seleção de Analisadores de Sentimentos Textuais

Os textos de cada um dos 1.219.482 de tweets selecionados até aqui são, então, classificados com o auxílio de algoritmo de análise de sentimentos. Por meio da identificação, extração, quantificação de estados afetivos e informações subjetivas, a análise de sentimentos tem a principal função de classificar a polaridade de um determinado texto, onde aspectos de uma entidade (ex: palavra, exclamação, emoji, etc.) podem ser positivas, negativas ou neutras [Ho et al. 2019, Lerman et al. 2018, Gallegos et al. 2016]. Dentre diversos analisadores de sentimentos disponíveis, selecionamos o VADER e o Senticnet, por serem largamente utilizados ao longo dos últimos anos pela comunidade científica e largamente validados [Ribeiro et al. 2016].

O VADER (Valence Aware Dictionary for Sentiment Reasoning) é um modelo para análise de sentimentos desenvolvido especialmente para o contexto de redes sociais, sem requerer treinamento, capaz de avaliar padrões como excesso de pontuações, utilização de letras maiúsculas, *emojis*, *emoticons* e conjunções que podem inverter a polaridade do sentimento da mensagem [Hutto and Gilbert 2014]. Comparado à 11 analisadores de sentimentos altamente recomendados como o LIWC (Linguistic Inquiry and Word Count), o Affective Norms for English Words (ANEW), e o SentiWordNet, além de outros analisadores de sentimentos baseados em aprendizado de máquina como o Naive Bayes, Maximum Entropy e Support Vector Machine(SVM), o VADER obteve uma performance comparável e, em muitos casos, até melhor na análise de sentimentos [Ribeiro et al. 2016]. Ao analisar entidades em um texto, o VADER retorna resultados em quatro categorias: *negative*, *neutral*, *positive* e *compound*, sendo este último uma normalização dos valores anteriores entre -1 (extremamente negativo) a 1 (extremamente positivo), 0 como neutro.

O SenticNet é um analisador de sentimentos intermediário entre redes neurais artificiais e sistemas simbólicos típicos. A base de conhecimento do SenticNet, disponibilizada em diversos idiomas, fornece um conjunto de 200.000 conceitos de linguagem natural, com definições semânticas, *sentics* e polaridades associadas [Cambria et al. 2014]. As semânticas definem informações denotativas associadas a palavras e expressões, e os *sentics* definem informações conotativas (valores de categorização de emoções) e as polaridades retornam resultados entre -1 (extremamente negativo) a 1 (extremamente positivo), tal como no analisador de sentimentos VADER. Mais especificamente, o SenticNet retorna os seguintes aspectos: *polarity_value* (resultado da polaridade (negativa ou positiva)); *polarity_intense* (resultado consolidado da polaridade considerando os resultados denonativos e conotativos); *mood_tags* (*tags* associadas à palavra); *sentics* (sentimentos denotativos) e *semantics* (significados semânticos para a palavra).

2.3. Tradução de textos

O VADER analisa sentimentos de textos em inglês e o SenticNet, apesar de possuir uma base de conhecimento em português, optamos por utilizar sua base em inglês por ser mais ampla. Por este motivo, foi necessário que traduzíssemos os tweets de português

para o inglês. Utilizamos o GoogleTrans [3], uma biblioteca disponível em Python que implementa a API do Google Translate [4] para a realização das traduções. Esta API é limitada por tempo e, para contornar esta limitação, cronometramos automaticamente os tempos de chamadas à API para que as requisições fossem realizadas com intervalos de 5 segundo entre elas.

O processo de tradução automática de textos de português para inglês pode auxiliar na correção de palavras e termos utilizados originalmente. Por outro lado, este mesmo processo de tradução pode eventualmente modificar o sentido do texto original, especialmente aqueles coletados de redes sociais onde as mensagens, em grande parte, utilizam abreviações, gírias e se abstém da utilização de acentuação e pontuação [Farias 2016]. Visando evitar traduções incorretas, fizemos a verificação manualmente de textos traduzidos, ao acaso, para observar o sentido dos textos traduzidos em relação aos originais e, embora houvessem alguns textos com sentido modificado, isto não prejudicou a análise de sentimentos processada posteriormente.

Um mesmo tweet pode ser "retweetado", ou seja, o mesmo texto criado por um usuário, contido em um tweet, pode ser encaminhado por outro usuário. Verificamos que 61% dos textos originais são retweetados e, durante o processo de tradução, verificamos se o texto do tweet já havia sido traduzido anteriormente. Esta verificação é feita a partir de um valor de *hash* calculado previamente para todos os textos dos tweets. Desta forma, esta verificação é feita pelo *hash* calculado e não pelo seu conteúdo. Esta estratégia possibilitou busca até 4 vezes mais rápidas nos testes realizados utilizando nossa base de tweets brasileiros.

3. Análise de Sentimentos de Tweets sobre a COVID-19

Os tweets selecionados tiveram a análise de sentimentos processada pelo VADER e pelo SenticNet, e as pontuações (*scores*) de sentimentos positivos, negativos e neutros obtidos são registradas na base de dados do MongoDB. Estas análises de sentimentos são processadas por meio de algoritmos criados em linguagem de programação Python, cujas bibliotecas para o VADER [5] e para o SenticNet [6] estão gratuitamente disponíveis.

Diferentemente do VADER, o SenticNet analisa sentimentos palavra-a-palavra, e não por textos completos. Por este motivo, precisamos pré-processar o texto, realizando a sua *tokenização* (separação em palavras), limpeza de urls, remoção de *stopwords* e caracteres inválidos. Estas ações foram realizadas por meio da biblioteca Spacy[7] do Python, especializado em processamento de linguagem natural, para que fossem submetidas somente palavras válidas ao SenticNet. O Spacy possui um modelo (*en_core_web_sm*), pré-treinado para a língua inglesa.

No SenticNet, o valor considerado como pontuação de sentimento do texto é calculado a partir da média dos valores de *polarity_intense* retornados para cada palavra analisada. O resultado somente é retornado pelo SenticNet caso a palavra analisada esteja presente na base de conhecimento utilizada. Por este motivo, adotamos como regra con-

[3] https://py-googletrans.readthedocs.io/en/latest/
[4] https://cloud.google.com/translate
[5] https://pypi.org/project/vaderSentiment/
[6] https://pypi.org/project/senticnet/
[7] https://spacy.io/

siderar os resultados somente de textos que tivessem, no mínimo, 50% de suas palavras com resultados, diminuindo a probabilidade de resultados discrepantes no pontuação do texto caso fossem analisadas uma ou poucas palavras. Pelas características do SenticNet, perde-se a capacidade de analisar pontuações ou caracteres comumente utilizados nas redes sociais, situações contempladas pela ferramenta VADER.

Ao final dos processamentos dos analisadores de sentimentos, obtivemos o total de 1.219.482 tweets analisados pelo VADER (a totalidade do conjunto de registros selecionados) e 448.782 tweets analisados pelo SenticNet. A Figura 1 apresenta a mediana e respectivas distribuições obtidas pelo VADER (esquerda) e SenticNet (direita). Observa-se que os resultados provenientes da utilização do SenticNet concentram a maioria das pontuações em valores neutros e ligeiramente positivos, enquanto o VADER concentra as pontuações dos tweets em valores negativos. O VADER considera elementos da linguagem informal, amplamente utilizados nas redes sociais, como repetição de pontuações e *emoticons*. Por considerar uma quantidade maior de tweets e mais opções na análise de textos dos tweets, optamos por escolher o VADER para as análises subsequentes deste trabalho, em detrimento do SenticNet.

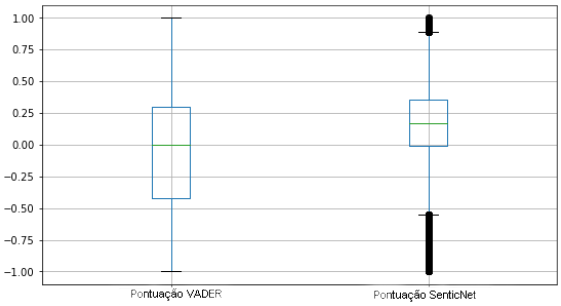

Figura 1. Distribuição de pontuações de análise de sentimentos obtidos pelo VADER (*boxplot* à esquerda) e SenticNet (*boxplot* à direita)

Após a escolha do VADER para análise de sentimentos, agrupamos os tweets por cidades e geramos, automaticamente, métricas de resultados acumulados para os textos. Para cada cidade, analisamos seus resultados em três diferentes períodos, dentro do intervalo de datas coletado: antes do primeiro caso de Covid-19 (pré-pandemia), após o primeiro caso de Covid-19 no Brasil (ocorrido em 26 de fevereiro de 2020 [8]) e após a primeira morte por Covid-19 no Brasil (ocorrida em 12 de março de 2020 [9]).

As quantidades totais de tweets e as quantidades por período foram calculadas somente para cidades que possuíssem dados nos três períodos analisados (pré-pandemia, primeiro caso e primeira morte por Covid-19 no Brasil), totalizando um conjunto com 890 cidades selecionadas. Ressalta-se que calculamos as médias das quantidades de tweets para cada período selecionando somente cidades que possuíssem, no mínimo, em cada intervalo, a quantidade média de tweets calculada do respectivo período analisado. Este critério foi utilizado para evitar discrepâncias nas médias de pontuações em cidades que

[8] https://coronavirus.saude.gov.br/linha-do-tempo/

[9] https://agenciabrasil.ebc.com.br/saude/noticia/2020-06/primeira-morte-por-covid-19-no-brasil-aconteceu-em-12-de-marco

possuíssem poucos tweets em dado intervalo. Com esta seleção, passamos a ter 86 cidades selecionadas, correspondendo a 984.807 tweets.

A partir deste conjunto de tweets para 86 cidades selecionadas, calculamos, para cada cidade, as médias totais de pontuações, as médias de pontuações nos três períodos analisados e geramos as entidades presentes nos tweets com maiores pontuações negativas e maiores pontuações positivas utilizando o Named Entity Recognition (NER), que consiste em classificar entidades (palavras e termos referentes a objetos, tais como nomes, lugares, organizações, etc.) de acordo com as classes definidas [Nadeau and Sekine 2007]. A utilização do NER possibilitou a identificação, como entidades, de termos formados por nomes e palavras compostas como "Jair Bolsonaro" e "Sistema Único de Saúde", ao invés de considerarmos cada palavra como um *token* individual.

O NER foi computado por meio da biblioteca Spacy, utilizando modelo *pt_core_news_sm* pré-treinado para a língua portuguesa. Este modelo possui as seguintes classes de entidades definidas: PER (nomes de pessoa ou família), LOC (localizações políticas ou geográficas), ORG (nomes de corporações, entidades governamentais, etc) e MISC (outras entidades como eventos, nacionalidades, produtos, etc).

A biblioteca Spacy também possui a função *Named Entity Ruler*, que possibilita a inclusão de padrões de reconhecimento de entidades por meio de regras. A partir da observação de *tokens*, bigramas e trigramas extraídos dos textos, com a utilização da biblioteca Natural Language Tool Kit (NLTK) [10], incluímos mais de 200 novos padrões ao modelo pré-treinado da biblioteca, contemplando nomes de personalidades públicas, partidos políticos, órgãos governamentais, veículos de comunicação, eventos, doenças, medicamentos, suprimentos hospitalares, entre outros. Este novos padrões permitiram o reconhecimento de diferentes formas de referência a um termo para que pudéssemos identificá-lo como uma entidade única, como "Jair Bolsonaro", "Bolsonaro" e "Presidente Bolsonaro" e a identificação de termos relevantes ao contexto da pandemia de Covid-19 como *lockdown* e *fake news*.

4. Resultados

Nas 86 cidades contendo tweets analisadas para o período total de 90 dias, obtivemos Q2 = -0.043985, desvio padrão de 0.015145 e intervalos de mínimo = -0.084883, Q1 = -0.053589, Q3 = -0.034656 e máximo = 0.004392. Nota-se que a maioria das pontuações concentram-se em 0, ou sentimento de neutralidade, não apresentando tendências significativas de sentimentos positivos ou negativos nas pontuações das cidades analisadas.

Em nova análise de resultados, demarcamos 3 períodos diferentes em um gráfico de séries temporais diárias, da seguinte forma: até a data do primeiro caso de Covid-19 ocorrido no Brasil em 26 de fevereiro de 2020 (barra vertical vermelha), período até a primeira morte por Covid-19 no Brasil em 12 de março de 2020 (barra vertical preta), e após a primeira morte. As 86 cidades estão incluídas nesta série temporal de 90 dias de coleta de dados do GeoCov19, como pode ser visualizado na Figura 2. Nota-se a estabilização das pontuações de sentimentos após o primeiro caso e a primeira morte por Covid-19 no Brasil, aparentando haver tendência de adaptação hedônica. A adaptação hedônica é um tendência visualizada em pessoas, onde verifica-se o retorno a um nível relativa-

[10] https://www.nltk.org/

mente estável emocional, apesar da ocorrência de importantes acontecimentos positivos ou negativos anteriormente [Diener et al. 2006].

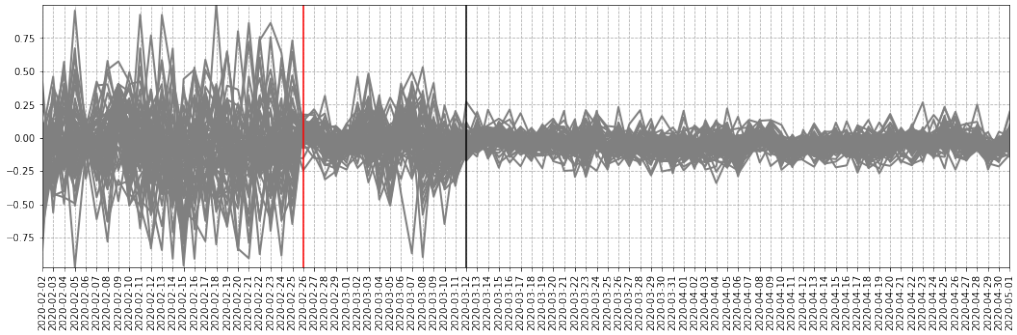

Figura 2. Médias diárias de 86 cidades brasileiras com pontuações de análises de sentimentos demarcados pela primeiro caso de Covid-19 (barra vermellha) e primeira morte (barra preta) no Brasil

As entidades, obtidas por meio do processamento de NER das 86 cidades analisadas, são ilustradas como nuvem de palavras na Figura 3. Verifica-se, nesta figura, as entidades com maior ocorrência contidas em todos os tweets analisados (a), entidades com maior ocorrência em tweets com maiores pontuações negativas (b), e de entidades em tweets com maiores pontuações positivas (c). Em todas as nuvens, as palavras mais destacadas acabam se repetindo, demonstrando repetição de tópicos ao longo dos 90 dias de dados coletados e analisados do GeoCov-19.

Figura 3. Nuvens de palavras, organizadas em 3 grupos: (a) por entidades existente em todos tweets analisados, (b) por entidades com maior ocorrência maiores pontuações negativas (b), e (c) por entidades com maiores pontuações positivas.

Inspirados nas demarcações (barras) do primeiro caso e da primeira morte por Covid-19 da Figura 2, fizemos uma análise das entidades obtidas por meio do processamento do NER, apresentado na Figura 4 e organizados em duas colunas: tweets com maiores pontuações negativas (esquerda, em vermelho), e tweets com maiores pontuações positivas (direita, em azul). Antes do primeiro caso no Brasil, a palavra ”China”, lugar do epicentro do Covid-19 neste período, ganha destaque nas duas colunas. Após o primeiro caso, os locais com maior quantidades de casos no Brasil (São Paulo) e no mundo (Estados Unidos) ficam evidenciados. Por último, após a primeira morte, evidencia-se o destaque ao nome do atual presidente do Brasil.

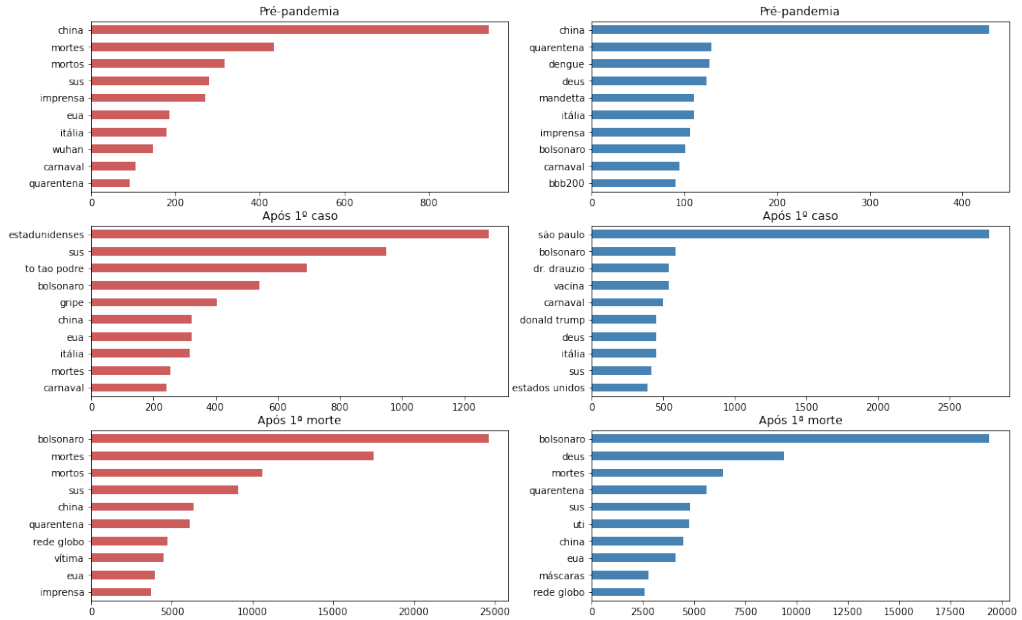

Figura 4. Ocorrências de entidades organizadas em duas colunas: tweets com maiores pontuações negativas (esquerda, em vermelho), e tweets com maiores pontuações positivas (direita, em azul)

5. Considerações Finais

Neste artigo, exploramos tendências positivas, negativas, e neutras de sentimentos compartilhados em tweets por pessoas ao longo de 90 dias, em períodos relevantes da pandemia de Covid-19 no Brasil, tais como o primeiro caso confirmado e a primeira morte. Realizamos a coleta destes tweets nos primeiros meses de 2020, o tratamento e a sua seleção além de comparar dois analisadores de sentimentos e aplicar aquele capaz de processar mais possibilidades de textos e maior quantidade de tweets, considerando que ambos analisadores são bem aceitos pela comunidade científica. Ao todo, foram 86 cidades brasileiras com tweets de pessoas analisadas, e identificamos 3 períodos relevantes: o primeiro caso de Covid-19 e da primeira morte por Covid-19 no Brasil, observando a estabilização dos scores de sentimentos após cada um destes marcos, evidenciando uma tendência de adaptação hedônica destes usuários de Twitter.

Alguns resultados obtidos correlacionando aspectos sociais e comportamentais em redes sociais estão sendo utilizados em políticas públicas e até mesmo como métricas em censos populacional [Plunz et al. 2019, Venerandi et al. 2015]. Neste sentido, os resultado comentados no presente artigo pode incentivar governantes e gestores públicos em novas iniciativas de políticas públicas, para amparar a população em períodos de maiores oscilações positivas e negativas de sentimentos, ou de reforço das medidas de proteção necessárias durante períodos de sentimentos neutros e de adaptação hedônica. A visualização destes resultados podem ser temporais e transversais, o que facilitará o monitoramento e a tomada de decisão por estes gestores.

Futuramente, pretendemos analisar outros dados disponíveis dos tweets coletados, como menções, retweets, além de relacionar a outros fatores que possam influenciar os textos compartilhados, como variáveis demográficas e socioeconômicas.

Agradecimentos

Os autores agradecem à Universidade de Fortaleza por patrocinar o projeto ”Elaboração de modelo fenomenológico preditivo para a maximização da liberação das atividades comerciais durante a pandemia de COVID-19 ”, que possibilitou o desenvolvimento do trabalho descrito neste artigo. O autor Mario Maia agradece esta mesma Universidade por prover a bolsa de incentivo à inovação e à pesquisa científica e tecnológica, a qual permitiu o fomento do seu trabalho de pesquisa e da escrita deste artigo.

Referências

Cambria, E., Olsher, D., and Rajagopal, D. (2014). Senticnet 3: A common and commonsense knowledge base for cognition-driven sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1).

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

de Melo, T. and Figueiredo, C. M. S. (2021). Comparing news articles and tweets about covid-19 in brazil: Sentiment analysis and topic modeling approach. *JMIR Public Health Surveill*, 7(2):e24585.

Diener, E., Lucas, R. E., and Scollon, C. N. (2006). Beyond the hedonic treadmill: revising the adaptation theory of well-being. *The American psychologist*, 61(4):305—314.

Farias, E. d. S. (2016). Relevância da tradução de textos de português para inglês no processo de classificação binária de sentimento de postagens rápidas em redes sociais on-line. Master's thesis, Universidade Federal de Campina Grande.

Gallegos, L., Lerman, K., Huang, A., and Garcia, D. (2016). Geography of emotion: Where in a city are people happier? In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 569–574.

Győrödi, C., Győrödi, R., Pecherle, G., and Olah, A. (2015). A comparative study: Mongodb vs. mysql. In *13th International Conference on Engineering of Modern Electric Systems (EMES)*, pages 1–6. IEEE.

Ho, V. A., Nguyen, D. H.-C., Nguyen, D. H., Thi-Van Pham, L., Nguyen, D.-V., Van Nguyen, K., and Nguyen, N. L.-T. (2019). Emotion recognition for vietnamese social media text. In *International Conference of the Pacific Association for Computational Linguistics*, pages 319–333. Springer.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1).

Jiang, J., Chen, E., Yan, S., Lerman, K., and Ferrara, E. (2020). Political polarization drives online conversations about covid-19 in the united states. *Human Behavior and Emerging Technologies*, 2(3):200–211.

Lerman, K., Marin, L. G., Arora, M., Lima, L. H. C., Ferrara, E., and Garcia, D. (2018). Language, demographics, emotions, and the structure of online social networks. *Journal of Computational Social Science*, 1(1):209–225.

Lopreite, M., Panzarasa, P., Puliga, M., and Riccaboni, M. (2021). Early warnings of covid-19 outbreaks across europe from social media. *Scientific Reports*, 11(2147).

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, 30(1):3–26.

Plunz, R. A., Zhou, Y., Vintimilla, M. I. C., Mckeown, K., Yu, T., Uguccioni, L., and Sutto, M. P. (2019). Twitter sentiment in new york city parks as measure of well-being. *Landscape and urban planning*, 189:235–246.

Qazi, U., Imran, M., and Ofli, F. (2020). Geocov19: A dataset of hundreds of millions of multilingual covid-19 tweets with location information. *ACM SIGSPATIAL Special*, 12(1):6–15.

Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., and Benevenuto, F. (2016). Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29.

Rodrigues de Andrade, F. M., Barreto, T. B., Herrera-Feligreras, A., Ugolini, A., and Lu, Y.-T. (2021). Twitter in brazil: Discourses on china in times of coronavirus. *Social Sciences Humanities Open*, 3(1):100118.

Venerandi, A., Quattrone, G., Capra, L., Quercia, D., and Saez-Trumper, D. (2015). Measuring urban deprivation from user generated content. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 254–264.

Xiang, Y.-T., Li, W., Zhang, Q., Jin, Y., Rao, W.-W., Zeng, L.-N., Lok, G. K. I., Chow, I. H. I., Cheung, T., and Hall, B. J. (2020). Timely research papers about covid-19 in china. *The Lancet (British edition)*, 395(10225):684–685.