UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Arthur Filipe Sousa Gomes

Um *framework* para geocodificação de tuítes do território brasileiro

Uberlândia, Brasil

2023

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Arthur Filipe Sousa Gomes

Um *framework* para geocodificação de tuítes do território brasileiro

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.

Orientador: Profa. Dra. Maria Camila Nardini Barioni

Universidade Federal de Uberlândia – UFU

Faculdade de Ciência da Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2023

Resumo

Quando se trata de pesquisas com dados de redes sociais, um dos desafios encontrados é a identificação da localização geográfica daqueles que, de alguma forma, interagem nesses ambientes virtuais. Nesse contexto, destaca-se o Twitter, que é uma das redes sociais mais utilizadas no mundo, com uma audiência estimada em mais de 368 milhões de usuários ativos por mês. Tendo em vista esse cenário, o presente trabalho abordou a tarefa de associar, de forma assertiva e automática, os dados do Twitter às localizações geográficas dos usuarios residentes no Brasil e que os originaram, o que é conhecido como geocodificação. Sendo assim, para embasar pesquisas com dados de redes sociais no âmbito brasileiro, objetivou-se a proposição de um *framework* para detecção da identificação da localização geográfica de usuários a partir de metadados do Twitter no Brasil e ainda, avaliar a acuracidade e confiança com relação à tal *framework*. Após a execução dos experimentos, considerou-se que os resultados obtidos foram satisfatórios porque obteve-se um Valor Preditivo Positivo (VPP) de 97,1% no teste final para detecção de localizações com granularidade em nível de município. Obteve-se também um VPP de 97,0%, quando se aplicou o *framework* em uma base de tuítes de trabalho correlato. O fato de que a solução criada alinhou-se com a justificativa apresentada, ao utilizar o parâmetro *Location* do campo *User* como geolocalizador, é outro detalhe positivo associado aos resultados. A análise interpretativa das nuvens de palavras, referentes aos tuítes geocodificados da base de tuítes do trabalho correlato, indicou que o conteúdo compartilhado no campo textual pelos usuários está relacionado com as regiões que eles compartilham no campo *Location*, o que trouxe confiança no uso do *framework* proposto. Além disso, demonstrou que, ao associá-lo com técnicas de análise de sentimentos, pode-se obter informações importantes sobre a reação da população brasileira diante de eventos importantes, como a pandemia da Covid-19, por exemplo.

Palavras-chave: Twitter, geocodificação, nuvens de palavras, *framework*, Covid-19.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

NP	Nuvem de Palavras
VPP	Valor Preditivo Positivo
CPI	Comissão Parlamentar de Inquérito
Qtd	Quantidade
FP	*False Positives*
TP	*True Positives*

Sumário

1 Introdução

Sabe-se que, com o aumento da proporção de pessoas com acesso à internet e a grande adesão às redes sociais nos últimos anos, tais espaços virtuais se tornaram o principal meio de comunicação e interação humana. Uma das consequências disso é o rápido compartilhamento de grandes quantidades de dados variados, que refletem as notícias e os assuntos mais discutidos do momento. Neste cenário, são criadas oportunidades, em diversas áreas, para estudos que possam vir a explorar esses dados, visando encontrar *insights* ou contribuir com pesquisas futuras. Áreas como Saúde, Administração e Gestão de Desastres são algumas das que podem ser beneficiadas com a análise de dados de redes sociais (HOU; HAN; CAI, 2020).

Quando se trata de pesquisas com dados de redes sociais, um dos desafios encontrados é a identificação da localização geográfica daqueles que, de alguma forma, interagem nesses ambientes virtuais. Isso porque, muitas vezes, é importante correlacionar algum aspecto dos dados com a localização geográfica daqueles responsáveis pela geração. Tal informação pode estar implícita, explícita, pode ser informada diretamente pelo usuário da rede social, de forma estruturada ou não. Pode ainda, nem ser informada. Nesse contexto, é possível utilizar métodos para detecção de localização dos usuários a partir de dados de redes sociais. Jiang e colaboradores, por exemplo, conduziram uma pesquisa em que um desses métodos foi utilizado em dados do Twitter, sendo que o intuito da pesquisa foi entender como as características políticas de alguns estados americanos poderiam influenciar na evolução das discussões online sobre a COVID-19 (JIANG et al., 2020).

Sobre o Twitter, sabe-se que essa é uma das redes sociais mais utilizadas no mundo, com uma audiência estimada em mais de 368 milhões de usuários ativos por mês, e que o formato de microblogging adotado pela rede possibilita que os usuários publiquem pequenos textos sobre o assunto que quiserem (DIXON, 2022). Isso permite que eventos da vida real sejam reportados quase que ao mesmo tempo em que acontecem. Nesse contexto, e durante a última década, a rede social tem sido utilizada por pesquisadores, por exemplo, na detecção de eventos, na análise de sentimentos do usuário e em muitas outras áreas de estudo (KARAMI et al., 2020). Enfim, por ser uma área a ser explorada, o presente trabalho abordará a tarefa de associar, de forma assertiva e automática, os dados do Twitter às localizações geográficas dos usuários que os originaram.

1.1 Objetivos

Ao se abordar o tópico métodos de detecção de localização a partir de dados de redes sociais, é possível destacar que ainda não se sabe qual é a acuracidade dos mesmos na detecção das localizações geográficas dos usuários das redes sociais, principalmente quando se pensa no contexto brasileiro. Com intenções semelhantes às do presente trabalho, Liu et al. (2021) conduziram uma pesquisa para avaliar a acurácia e a robustez dos métodos com dados de *check-in* do Instagram em Hong Kong . Nesse mesmo sentido, e para embasar pesquisas com dados de redes sociais no âmbito brasileiro, o presente trabalho tem o objetivo de propor uma metodologia para detecção da localização geográfica de usuários a partir de metadados do Twitter no Brasil e ainda, avaliar a acuracidade do método de detecção automática da localização a ser proposto.

1.2 Justificativa

A pesquisa proposta aqui é importante para apoiar futuros estudos que possam utilizar dados do Twitter no Brasil, independente da área científica a ser explorada. Sabe-se que, nem todos os usuários ativos nas redes sociais informam as suas localizações geográficas em suas interações e, além disso, tal informação pode não aparecer de forma estruturada ou explícita. Tendo em vista tal situação, verifica-se necessário o desenvolvimento de um método de avaliação dos dados do Twitter eficaz e confiável na detecção de localização individual do usuário. Nesse sentido, estudos que pretendem avaliar a correlação geográfica com aspectos da realidade serão os beneficiados.

No contexto proposto até aqui, deve-se levar em conta a possível indisponibilidade de informação nos campos que indicariam a localização real do usuário nos metadados do Twitter, como por exemplo, *Latitude*, *Longitude* e *Place*. Ademais, a possibilidade do uso do parâmetro *Location* do campo *User*, que é de livre preenchimento do usuário, como geolocalizador é também um motivador importante desta pesquisa. Para exemplificar o problema e fortalecer a justificativa, foi realizada uma hidratação de tuítes por meio da aplicação *Hydrator*, ou seja, buscou-se os metadados a partir dos IDs dos tuítes, em duas bases de dados disponibilizadas para o público geral para análise dos campos já citados (NOW, 2022). Na primeira, uma base coletada pela UFABC, sobre a CPI da COVID, entre as datas 25/10/2021 e 26/10/2021, em uma amostra de 50 mil tuítes, apenas 400 apresentaram informação no campo *Place*. Já o campo *User.Location* apresentou conteúdo em todos os 50 mil tuítes (UFABC, 2021). Na segunda, disponibilizada online por Tiago de Melo, da Universidade do Estado do Amazonas, em uma amostra de também 50 mil tuítes, apenas 1.121 tuítes possuem o campo Place preenchido, enquanto que o *User.Location* apresentou conteúdo em 19.811 tuítes (MELO, 2020). Devido a essa maior disponibilidade de informação no campo *Location*, torna-se mais viável propor uma metodologia que utiliza

tal parâmetro para geocodificação dos tuítes e, ao mesmo tempo, avaliar se a mesma é segura ou se pode trazer incertezas capazes de prejudicar os resultados de futuras pesquisas que envolvam dados de usuários do território brasileiro.

1.3 Organização do Texto

O texto deste trabalho é composto pelo capítulo 2, Análise de Dados de Redes Sociais, cuja proposta é apresentar, de forma geral, como são realizadas as pesquisas e análises com dados de redes sociais, com destaque para os conceitos abordados ao longo do texto e contextualização com a pesquisa aqui realizada. O capítulo seguinte, Trabalhos Correlatos, aborda pesquisas de outros autores que também envolveram alguns dos temas relatados aqui. Em seguida, no quarto capítulo, descreve-se o método de trabalho desta pesquisa, adotado com o intuito de elaborar o *framework* previsto e possibilitar sua aplicação em uma base de dados de tuítes, relacionada à pandemia da COVID-19. O capítulo Resultados e Discussões (quinto capítulo) apresenta, primeiramente, todos os resultados obtidos referentes aos testes para a definição do *framework*, referentes à aplicação do *framework* na base mencionada e referentes à geração de nuvens de palavras a partir do conteúdo do texto dos tuítes da mesma base. Em seguida, apresenta toda a discussão de tais resultados, juntamente com a análise de sentimentos realizada por meio da análise interpretativa das nuvens de palavras. Finalmente, o último capítulo conclui o trabalho com os novos conhecimentos adquiridos e, além disso, dá recomendações para a continuidade desta pesquisa.

2 Análise de Dados de Redes Sociais

O presente trabalho aborda temas relacionados à análise de dados de redes sociais. Tal analise permite a descoberta de informações importantes sobre determinado conjunto de dados. Pode-se, por exemplo, fazer a análise textual de publicações por meio de algoritmos de classificação e apontar uma localização geográfica para os usuários que geraram tais dados. Trata-se, nesse caso, de uma maneira de se realizar a geocodificação dos dados. Essas fontes de informações estão cada vez mais fáceis de se adquirir e se apresentam em formatos simples nos metadados das redes sociais. Os conceitos citados aqui serão abordados ao longo do trabalho e, dessa forma, torna-se necessário elucidá-los.

Diferente dos conjuntos de dados tradicionais, cujas fontes produtoras são bem definidas, os dados de redes sociais podem ser criados por quaisquer usuários e, geralmente, estão em linguagem natural. Além disso, com a grande quantidade de usuários, tais dados são criados em grandes quantidades, a uma velocidade alta e apresentam grande variedade. Outro aspecto é que não existem controles de qualidade para os mesmos, o que faz com que a informação criada não tenha tanta credibilidade e integridade como a de fontes oficiais. Nesse sentido, muitas pesquisas são realizadas com o intuito de enfrentar tais dificuldades e gerar informações importantes a partir dos dados (HOU; HAN; CAI, 2020).

Em geral, o processo completo de análise de dados de redes sociais é composto pelas seguintes etapas: coleta de dados, limpeza e armazenamento, análise de dados e visualização de resultados (HOU; HAN; CAI, 2020). Tendo em vista tais etapas e o método de trabalho utilizado na presente pesquisa, elaborou-se uma revisão bibliográfica abordando cada uma delas, associando-as às técnicas utilizadas em cada parte do método de trabalho.

2.1 Coleta de Dados

Trata-se da tarefa de obter os dados das fontes, que no caso, são as redes sociais. Atualmente, a maioria dessas redes dispõem de APIs, *application programming interface (API)* interface de programação de aplicativo (ou aplicação) , um software intermediário que possibilita a comunicação entre dois aplicativos, para que quaisquer interessados realizem a coleta dos dados. Além disso, pode-se coletar dados de algum conjunto de dados públicos (*public dataset*) oficiais disponíveis na web (MARTINS et al., 2020). Em alguns casos, como no do Twitter, é possível realizar a coleta de dados em tempo real com bom desempenho (HOU; HAN; CAI, 2020).

Tabela 1 – Exemplo de tuíte com os metadados principais

Campo	Conteúdo
created_at	Tue Apr 27 23:57:37 +0000 2021
id	1387194251772104706
full_text	O Zero Um quer o lockdown da CPI, só da CPI. #CPIdaCovid #FlavioBolsonaro #COVID19 #SenadoFederal https://t.co/EsYM4dKNKi
source	Twitter for iPad
coordinates	
place	
lang	pt
metadata.result_type	recent
user.id	3011101
user.name	Marinhos
user.screen_name	marinhos
user.location	Here, there and everywhere.
user.description	Educação, tecnologia, ciência, política. Com algum humor. Education, technologies, science, politics. With some humor.
user.protected	False
user.followers_count	447
user.friends_count	322
user.listed_count	18
user.created_at	Fri Mar 30 20:27:11 +0000 2007
user.favourites_count	34
user.geo_enabled	False

2.1.1 Metadados do Twitter

Como o Twitter é a rede social envolvida no estudo, é importante saber como são estruturados os dados de suas publicações. Sabe-se que suas publicações são acompanhadas por um arquivo *Javascript Object Notation (JSON)*. Tal arquivo pode conter os metadados da publicação, compostos pelo perfil do usuário, data de registro do perfil, localização do usuário, número de seguidores e amigos, coordenadas GPS e horário em que o tuíte foi postado. No entanto, uma porção pequena dessas mensagens apresenta as coordenadas GPS que indicariam a localização exata do usuário durante a publicação, pois não é um campo obrigatório. Sendo assim, os métodos de classificação e de detecção de localização, existentes ou que ainda serão desenvolvidos, podem ser utilizados como forma de indicar a localização geográfica dos usuários, a partir daquilo que foi compartilhado pelos mesmos nos textos das publicações (ZHANG; GELERNTER, 2014). A Tabela 1 traz um exemplo de tuíte obtido para o presente trabalho, com 20 dos 179 campos disponíveis para análise de cada tuíte.

2.2 Limpeza e Armazenamento

Esta seção diz respeito ao processo de preparar a base de dados para o processo de análise e à definição da melhor maneira de armazenar tais dados, respectivamente (MARTINS et al., 2020).

2.2.1 Pré-Processamento de Dados Textuais

Conforme já mencionado, dados oriundos de redes sociais são, em algumas situações, dados em linguagem natural que muitas vezes não passam por nenhum controle de qualidade e, dessa forma, podem apresentar características pouco importantes para um processo de análise de dados. Por essa razão, é preciso conhecer cada vez mais sobre o processamento de linguagem natural (PLN), que é a área voltada a estudar a forma como as máquinas e computadores podem interpretar e responder utilizando a linguagem natural. Considerando a PLN, sabe-se que, caso não haja uma limpeza de dados, pode-se criar barreiras que dificultam o entendimento da linguagem dos seres humanos por parte da máquina. Sendo assim, o pré-processamento de texto é recomendado para a execução de atividades de PLN, já que traduz o texto bruto para uma versão mais "limpa", o que permite que as máquinas consigam entender, interpretar e trabalhar de maneira otimizada (MARTINS et al., 2020; TIBURCIO, 2021).

De acordo com (MARTINS et al., 2020), o pré-processamento de texto apresenta as seguintes etapas principais: limpeza de caracteres especiais, retirada de *stopwords*, *stemming* e lematização. As etapas são explicadas a seguir, com exceção da limpeza de caracteres especiais, cujo nome é autoexplicativo.

- Remoção de *stopwords*:

  Trata-se de um método que remove palavras muito frequentes de um dado textual, já que, na maior parte das vezes, as mesmas não se traduzem em informações muito relevantes para o entendimento do texto. Esse processamento pode se aplicar a qualquer linguagem, no entanto, pode não ser interessante realizá-lo, dependendo do contexto de análise. Na análise de sentimentos, por exemplo, em que se verifica se o sentimento sobre um assunto é positivo ou negativo, não é tão interessante remover a *stopword* "não", uma vez que traz uma conotação de negatividade para a frase, apontando justamente o sentimento transmitido. Diversas listas de *stopwords* da língua portuguesa podem ser encontradas na internet. (MARTINS et al., 2020)

- Lematização:

  Trata-se do processo de converter as palavras nos lemas correspondentes, com o objetivo de agrupar diferentes palavras que possuem uma forma base em comum: a forma no masculino e no singular. Por exemplo, palavras como gato, gatas, gata e gatos são reduzidas ao lema gato (MARTINS et al., 2020; TIBURCIO, 2021).

- *Stemming*:

  O *Stemming* é diferente da Lematização de maneira sutil, já que se trata da redução de uma palavra para o seu radical, como as palavras meninas e menino, que se reduziriam ao radical menin. Assim como a remoção das *stopwords*, nem sempre

este processamento e o anterior são indicados, dependendo da aplicação, pois podem remover informações textuais importantes (MARTINS et al., 2020; TIBURCIO, 2021).

2.2.2 Armazenamento de dados

Para armazenamento dos dados de redes sociais em pesquisas atuais, geralmente são utilizados bancos de dados relacionais (como o Microsoft SQL Server e MySQL) e bancos de dados não-relacionais (como o Cassandra e Redis). Para esse último, tem-se percebido uma maior utilização, porém ainda existe a desvantagem de que, em algumas soluções NoSQL, ainda não é possível a realização de consultas (HOU; HAN; CAI, 2020).

2.3 Análise de Dados

Tendo em vista o contexto apresentado e, para se realizar a análise dos dados ao longo dessa pesquisa, fez-se necessário o desenvolvimento de um algoritmo de classificação. De acordo com Tan, Steinbach e Kumar (2009), a classificação é o processo de utilizar uma função f que mapeia um conjunto de atributos x em um conjunto de variáveis predefinidas y, chamadas de rótulos de classe. Conhece-se a função f também, de maneira informal, como modelo de classificação. Tal modelo pode gerar uma informação descritiva, ou seja, é capaz de servir como ferramenta explicativa para diferenciar objetos ou classes diferentes, mas também pode realizar a previsão de rótulos de classe para registros desconhecidos. Portanto, algoritmos de classificação são, resumidamente, técnicas de mineração de dados que envolvem a atribuição de um rótulo aos dados de entrada não classificados.

Conforme mencionado, uma das principais características das técnicas de classificação é o fato de mapearem conjuntos de dados não rotulados em classes pré-definidas (ASSIS, 2018 apud KOHAVI, 1995), sendo que, para fazerem isso, tais técnicas constroem modelos de classificação a partir de dados de treino, que possuem informação quanto às classes que pertencem. Após isso, pode-se aplicar tais modelos em bases de teste, em que nenhuma informação a respeito das classes é disposta, com o intuito de catalogar essas bases com relação aos rótulos previamente conhecidos (ASSIS, 2018 apud HAN; KAMBER, 2011).

Como exemplos desses algoritmos, pode-se citar as Árvores de Decisão, o Classificador de Naive Bayes e o *Support Vector Machine* (MUKHERJEE; SAHANA; MAHANTI, 2017). Esses e outros exemplos mais conhecidos já estão implementados em bibliotecas de linguagens de programação, como na linguagem Python, por exemplo, que possui a Scikit-learn (PEDREGOSA et al., 2011). Ao aplicar tais técnicas em dados textuais do Twitter, por exemplo, pode-se encontrar padrões ou, em casos específicos como o dessa pesquisa, textos que indicam alguma localização geográfica.

2.3.1 Geocodificação de Dados

Um conceito importante para se apresentar é o de geocodificação, pois é exatamente o que se pretendeu realizar aqui. Trata-se do processo de traduzir expressões textuais de localização em um texto para a correta localização física. Esse é um processo que, todavia, apresenta complexidades, uma vez que existem lugares que possuem a mesma denominação, por exemplo, o que gera ambiguidades e incertezas. A corretude dessa indicação é outro obstáculo, já que, muitas vezes, há pouco contexto em uma mensagem de poucos caracteres ou muita "sujeira" para "limpar", como em campos textuais do Twitter (ZHANG; GELERNTER, 2014).

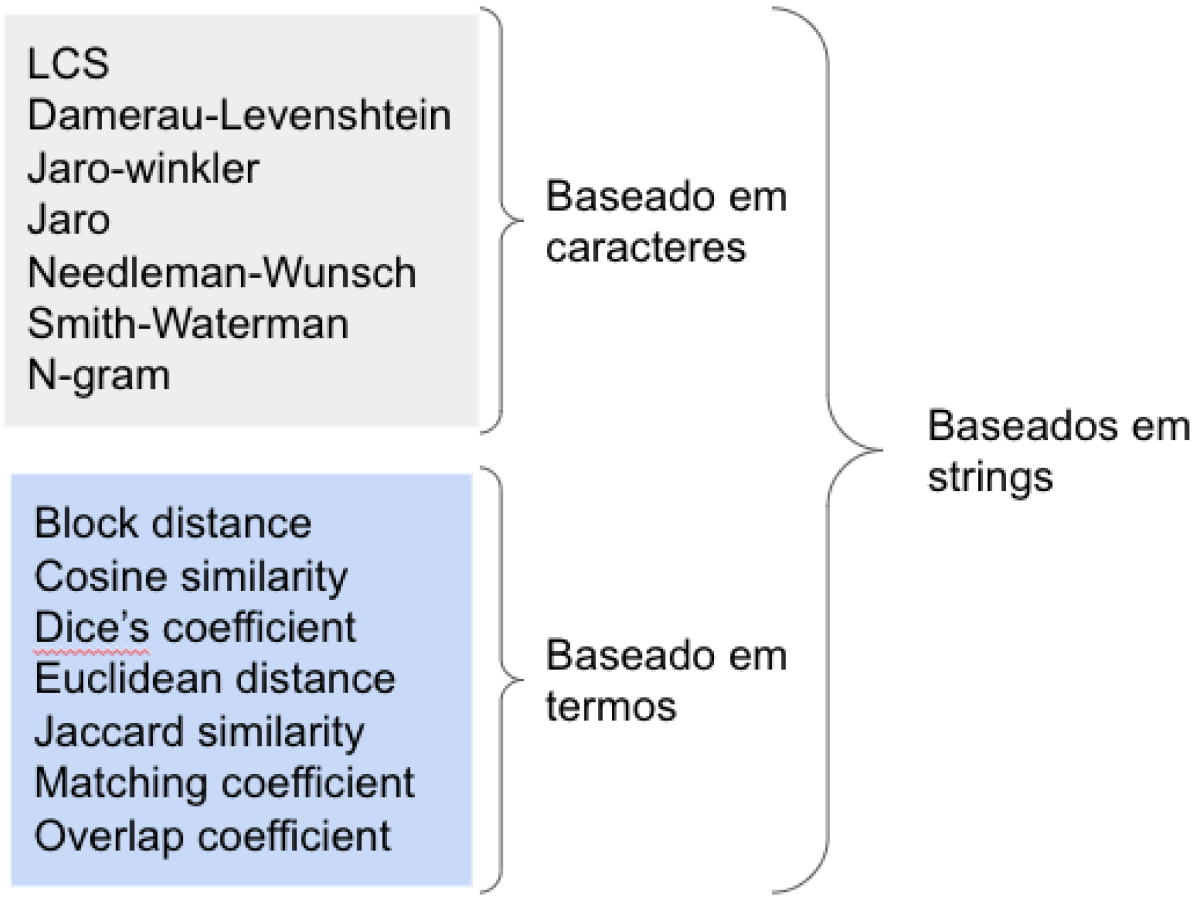
2.3.2 Similaridade Textual

Outro conceito importante na área de processamento de linguagem natural e para a presente pesquisa é a Similaridade Textual. Ao "calcular" tal similaridade, pode-se dizer que se pretende determinar a proximidade ou o quão dois fragmentos de texto são similares. Esse processo pode ser aplicado em diversos contextos como: classificação de texto, clusterização, sumarização de texto, detecção de tópicos e etc. A similaridade textual apresenta duas classificações que são a léxica e a semântica. A principal diferença entre as duas é que a similaridade semântica leva em consideração o significado das palavras ou da frase no contexto, enquanto a similaridade léxica diz respeito à similaridade estrutural do texto (MARTINS et al., 2020).

Com relação à similaridade léxica, pode-se determiná-la em diferentes níveis de granularidade: em nível de caractere, de palavra ou de frase. Normalmente, a similaridade em nível de caracteres é usada para determinar a proximidade de duas *strings* por meio de seus caracteres. Em geral, para se determinar tudo isso, utilizam-se métricas e métodos encontrados na literatura, que na sua essência, envolvem a verificação do número de operações necessárias para transformar uma *string* na outra (MARTINS et al., 2020).

Gomaa e Fahmy (2013) conseguiram listar 14 das principais abordagens criadas para cálculo da similaridade léxica (Figura 1), sendo que cada uma pode ser encaixada em uma das duas categorias em que se pode classificar os algoritmos: baseado em caracteres e baseado em termos. Uma vez que cada algoritmo funciona melhor em contextos diferentes, é mais interessante aprofundar o entendimento em relação àquele que foi utilizado neste estudo (MARTINS et al., 2020).

Figura 1 – Algoritmos baseados em *strings*.



Fonte: Adaptada de Gomaa e Fahmy (2013)

Levando-se em consideração a presente pesquisa, que objetivou criar um método para geocodificar tuítes em nível municipal, utilizando um campo em que o usuário geralmente preenche o nome da cidade em que reside, a similaridade semântica, ou de significado, não é um fator tão importante quanto a léxica. Isso porque, a biblioteca *Fuzzy Wuzzy*, utilizada para comparar *strings* e classificar os dados dos tuítes neste trabalho, se baseia na Distância de Levenshtein que, por sua vez, se apoia na similaridade léxica (FOUNDATION, 2022).

A Distância de Levenshtein considera que a similaridade entre duas *strings* é definida por meio do número mínimo de operações de edição (operações como inserção, exclusão e substituição) necessárias para transformar uma *string* em outra. Nesse cenário, operações de inserção e deleção apresentam custo igual a 1, mas a operação de substituição tem custo igual a 2, pois tal método considera a operação de substituição como a soma de uma operação de deleção e uma operação de inserção. Então, por exemplo, para transformar uma string1 ABADAC na string2 CADA, haveria o custo total de 4 operações (MARTINS et al., 2020).

1 operação de exclusão do A inicial (custo 1) + 1 operação de substituição da letra B por C com custo 2 (considerada uma inserção somada a uma exclusão) + 1 operação de exclusão da letra C no final da string (custo 1) = 4.

2.3.3 Uso da biblioteca *Fuzzy Wuzzy*

Trata-se de uma das maneiras mais fáceis de se fazer comparação textual com a linguagem de programação Python, uma vez que informa, com um valor de 0 a 100%, o nível de similaridade entre dois fragmentos de texto. Desenvolvida por *SeatGeeks*, tais características permitem o uso dessa funcionalidade em diversas aplicações de Processamento de Linguagem Natural. Pode-se, por exemplo, avaliar a performance de ferramentas de resumo de textos ou avaliar a autenticidade de textos. O fato de tal método retornar um score ou nota de similaridade permite a comparação relativa entre resultados, ou seja, dá embasamento para pesquisas que envolvam análise de dados (MAJUMDER, 2021). Basicamente, a biblioteca dispõe de várias funções e módulos que se traduzem em maneiras diferentes de se realizar a comparação entre *strings*. Tal comparação, em situações com dados reais, geralmente se resume em encontrar a *string* mais similar, dentre uma lista de opções, a uma *string* pré-determinada. Felizmente, o módulo *process* da biblioteca permite esse processamento, como o próprio nome já indica. A Figura 2 mostra um exemplo em que se pretende encontrar a *string* mais similar à marca de óculos "Ray-Ban", dentro de uma lista de *strings*. Na mesma figura, percebe-se o uso da função *process.extract* com os parâmetros "marca_pre_determinada", que recebe "Ray-Ban", "opcoes", que recebe a lista onde deve-se pesquisar as similaridades e limit, que define o número de tuplas que a função deve retornar, considerando as três maiores similaridade encontradas. Ao se analisar o retorno na Figura 2, nota-se que as *strings* "Ray", "rai ban" e "Chilli Beans", foram as *strings* que o algoritmo da *Fuzzy Wuzzy* considerou mais similares a "Ray-Ban", com 90%, 86% e 51% de similaridade, respectivamente (TUYCHIEV, 2020).

Figura 2 – Exemplo de uso do módulo process com resultado.

```
1 from fuzzywuzzy import process
2 marca_pre_determinada = 'Ray-Ban'
3 opcoes = ['Oakley', 'rai ban', 'Ray', 'Prada', 'Chilli Beans']
4 process.extract(query=marca_pre_determinada, choices=opcoes, limit=3)
```

```
[('Ray', 90), ('rai ban', 86), ('Chilli Beans', 51)]
```

Fonte: autoria própria

Ainda nesse contexto, pode-se destacar que existem cinco métodos de comparação de *string* disponíveis na função *process.extract*. São eles: *fuzz.ratio*, *fuzz.partial_ratio*, *fuzz.token_sort_ratio*, *fuzz.token_set_ratio* e *WRatio*. Por padrão, a função utiliza o método *WRatio* (*Weighted Ratio*), para definir as maiores similaridades, pois, de acordo

com Tuychiev (2020), esse é o método mais flexível, é indicado para o uso geral e retorna os melhores resultados.

2.3.4 *Frameworks*

De acordo com o dicionário, a palavra *framework* pode ser traduzida como uma armação, estrutura ou esqueleto. Pode ainda significar um sistema de ideias, regras ou crenças (INFOPEDIA, 2003-2021). Quando aplicada no contexto científico, essa palavra ganha um significado a mais. Passa a representar um guia ou passo-a-passo para realizar determinada tarefa ou alcançar algum objetivo.

2.3.5 Técnicas primárias de análise de dados

Entre as técnicas primárias, tem-se: a Análise de Assuntos. que consiste na adocao de métodos para a extração de assuntos ou informações significantes dos dados das redes sociais, os métodos de Análise de Sentimentos, que tentam, a partir dos dados, chegar a alguma conclusão quanto ao sentimento dos usuários frente a determinado assunto. Tem-se também, a Análise Temporal, que é utilizada para encontrar *insights* ou diferenças nos dados ao longo do tempo. Por fim, a Análise de Rede, se dedica a estudar as relações que ocorrem na comunidade virtual e entender como a informação flui e é disseminada (HOU; HAN; CAI, 2020).

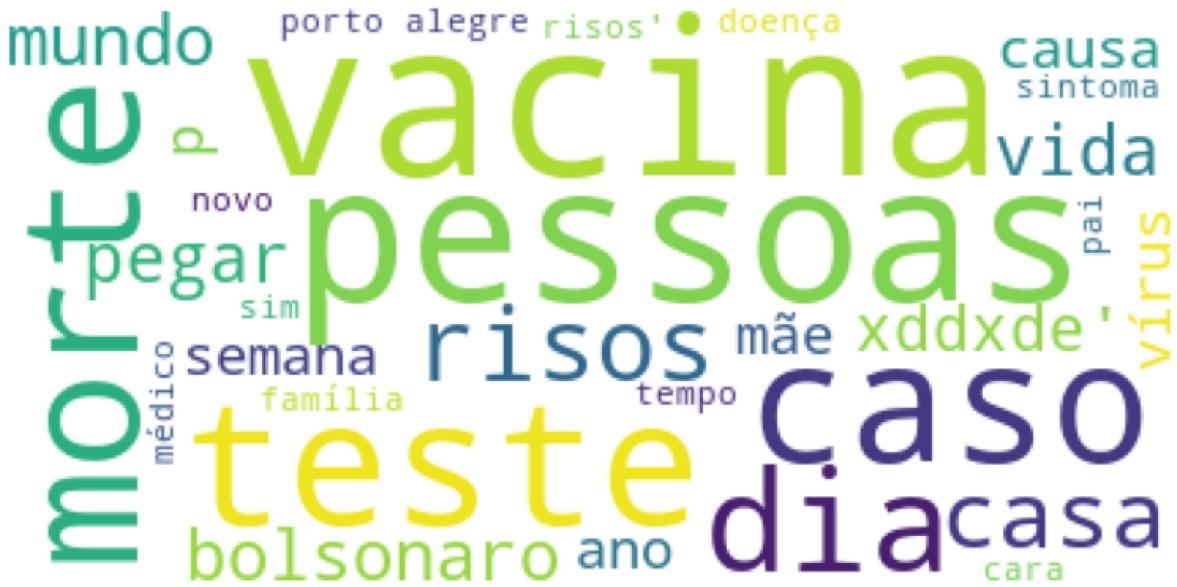
Uma vez que o presente trabalho foca na Análise de Sentimentos, faz-se necessária uma melhor contextualização com relação a esse tópico. De acordo com Tardelli, Dias e França (2019 apud REIS et al., 2015), uma das etapas para se realizar a análise de sentimentos é definir técnicas automáticas capazes de gerar informações subjetivas de textos em linguagem natural, como opiniões e sentimentos, a fim de criar conhecimento estruturado. Isso pode ser alcançado por meio de técnicas de visualização de dados, conforme discussão a seguir.

2.4 Visualização de Dados

A etapa de visualização de dados vem para ajudar quanto à definição mencionada no tópico anterior, uma vez que se trata da utilização de gráficos, tabelas ou outros tipos de visões como forma de representação dos dados. Visualizações bem desenhadas permitem e facilitam a descoberta de relacionamentos, tendências e padrões nos dados analisados, principalmente quando se tratam de grandes volumes de dados, como os de redes sociais. Dentre as técnicas automáticas existentes, destacam-se aquelas capazes de representar os dados obtidos através de imagens, de maneira a mostrar novas possibilidades de interação e compreensão dos mesmos. Nesse cenário, pode-se introduzir o conceito de *word clouds* (Figura 3), ou nuvens de palavras, que utiliza formas de percepção visual para facilitar a

compreensão dos termos mais relevantes, de forma generalizada, presentes em um texto geralmente muito grande (TARDELLI; DIAS; FRANÇA, 2019). Nas *word clouds*, faz-se o uso de recursos gráficos, como a mudança do tamanho da fonte de um texto que possui maior relevância ou possui um número maior de repetições na base analisada, algo que facilita bastante as análises interpretativas dos dados (TARDELLI; DIAS; FRANÇA, 2019).

Figura 3 – Exemplo de nuvem de palavras.



Fonte: autoria própria

3 Trabalhos Correlatos

Neste capítulo são apresentados alguns trabalhos que utilizaram algumas das técnicas de análise de dados conceituadas anteriormente.

3.1 Geocodificação

A utilização e criação de métodos relacionados às análises supracitadas têm crescido ao longo dos anos, juntamente com a popularização das redes sociais. Aliado a isso, também cresceu a necessidade de se desenvolver técnicas de análise de dados de redes sociais mais assertivas e confiáveis. Cresceu ainda, a necessidade de se verificar a correlação espacial dos assuntos mais discutidos nas redes, uma vez que podem trazer *insights* importantes sobre, por exemplo, a disseminação de uma doença em diferentes regiões ou a ocorrência de desastres naturais. Essa inserção da variável espacial na análise de dados de redes sociais acaba por trazer novos desafios, sendo um deles, a escassez de informação referente à geolocalização, uma vez que não é obrigatório informá-la nos campos próprios para tal (MOREIRA; BAKLIZKY; DIGIAMPIETRI, 2018). Como consequência, a comunidade científica tem se dedicado a encontrar as melhores maneiras de coletar tal informação e com a melhor precisão possível.

Moreira, Baklizky e Digiampietri (2018), por exemplo, utilizaram técnicas baseadas em frequência de palavras, algoritmos de classificação, como a Rede Bayesiana, e uma combinação das duas estratégias para indicar se postagens em Língua Portuguesa no Facebook possuíam ou não referência a uma localidade, bem como para comparar a acurácia dos resultados. Saldana-Perez et al. (2019), por sua vez, citam métodos que utilizam dicionários geográficos para estimar as coordenadas geográficas de um tuíte, baseando-se em seu corpo textual. Trata-se de um processo de 3 etapas que são: identificação dos elementos geográficos como ruas ou avenidas no corpo textual, seguida de uma busca pelas coordenadas no dicionário e, enfim, o registro da coordenada do tuíte, por meio da aplicação de funções geoespaciais nas coordenadas encontradas no passo anterior (SALDANA-PEREZ et al., 2019).

Para realizar a correlação de tuítes sobre a pandemia do COVID-19 e a localização dos usuários responsáveis pelas postagens nos Estados Unidos, ou seja, postagens em inglês, Jiang et al. (2020) utilizaram um algoritmo de correspondência textual denominado *Fast Fuzzy*. O objetivo em si foi detectar os nomes dos estados ou a abreviação dos mesmos nos dados. Além disso, realizaram uma verificação manual em uma amostra dos dados do Twitter que possuía informações geoespaciais para atestar a acuracidade do algoritmo, sendo que foram encontradas acuracidades altas para as localizações sugeridas

(JIANG et al., 2020). Pode-se perceber, portanto, que são diversas as abordagens para a detecção ou inferência da localização de usuários por meio de suas postagens ou mensagens compartilhadas em redes sociais. Todavia, a informação sobre a acurácia dos métodos costuma ser muito pontual e específica para os casos dos estudos aqui abordados. Sendo assim, torna-se importante construir um *framework* simples com acurácia satisfatória que indique ou classifique características geoespaciais das interações de usuários nas redes sociais. Além disso, quando se pensa na característica de microblogging do Twitter e a adesão dessa rede social no Brasil, a correlação do conteudo compartilhado com a localização dos usuários pode indicar importantes *insights* sobre o país e sua população.

3.2 Representação de dados com nuvens de palavras e análise interpretativa

Para demonstrar a eficácia das nuvens de palavras como instrumento para a representação de dados, pode-se citar o trabalho de Banni et al. (2020), que utilizaram desse recurso para realizar uma análise interpretativa detalhada, com o objetivo de identificar as alterações de comportamento e demandas relacionadas à religião e espiritualidade causadas pela pandemia do COVID-19. Com a análise das nuvens geradas antes e durante a pandemia, tais pesquisadores foram capazes de apontar até mesmo mudanças relativas nos sentimentos de usuários de diferentes religiões. Em outras palavras, pode-se considerar que as nuvens de palavras são um recurso simples, porém poderoso para a análise de sentimentos, principalmente quando se tem um grande volume de dados.

3.3 *Frameworks*

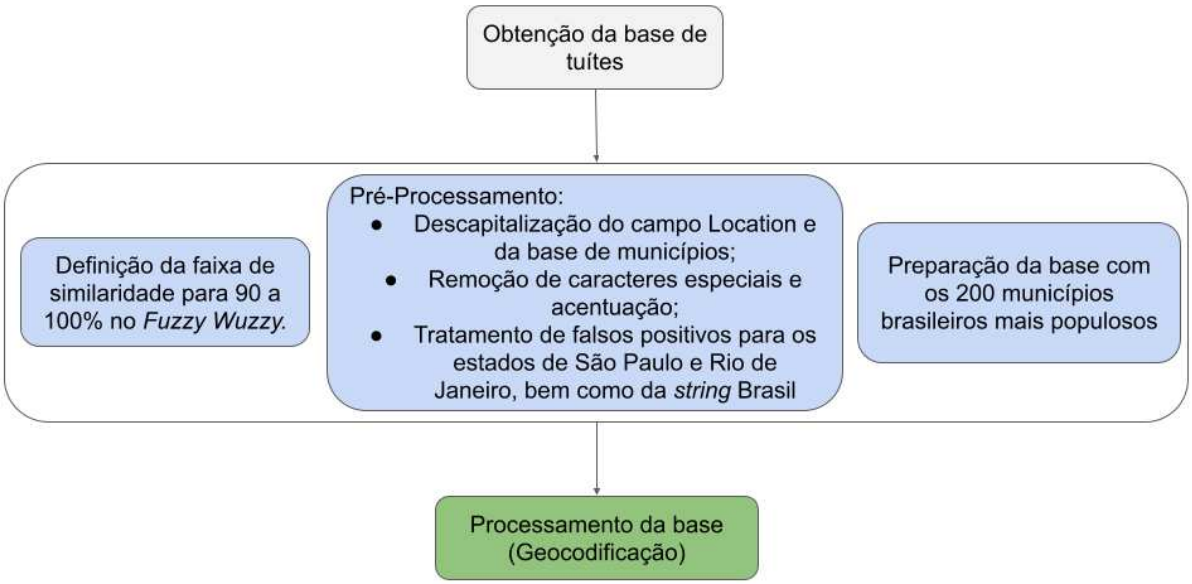
Frameworks tem sido utilizados em diferentes áreas, mas ganham destaque na análise de dados, podendo ganhar características bastante específicas dependendo da aplicação. Por meio dessa ferramenta, pode-se, por exemplo, detectar usuários com início de depressão a partir de dados de redes sociais, detectar tuítes relacionados à gripe ou ainda, realizar a classificação de textos árabes (ZRIGUI et al., 2012; ALESSA; FAEZIPOUR; ALHASSAN, 2018).

Nos exemplos de aplicação de *framework* para análise de dados de redes sociais citados, uma das etapas do *framework* consistiu na escolha de uma técnica analítica de dados primária. Tais técnicas são utilizadas na maioria dos estudos, dependendo daquilo que se deseja analisar, sendo que, pode-se listar quatro tipos principais: Análise de Assuntos, Análise Temporal, Análise de Sentimentos e Análise de Rede (ou da comunidade virtual) (HOU; HAN; CAI, 2020).

4 Método de Trabalho

O método a ser detalhado a seguir, relata como foi o processo de criação do *framework* para geodificação de tuítes brasileiros presente na Figura 4, que vai desde a obtenção dos dados, passa pelos testes para se chegar à faixa de similaridade de *strings* ótima a ser empregada e termina com a identificação de oportunidades para melhoria do *framework*. Por fim, aborda-se a aplicação do *framework* na base 2 (Tabela 2), com tuítes coletados durante a pandemia da COVID-19, o que possibilitou a análise de sentimentos a partir das nuvens de palavras geradas com o conteúdo dos tuítes, após esses terem sido georreferenciados pelo *framework*.

Figura 4 – *Framework* criado a partir dos resultados dos testes de similaridade.



Fonte: autoria própria

4.1 Obtenção da base de dados

O experimento se iniciou com a coleta de metadados da rede social Twitter, que foram extraídos em formato JSON e convertidos em formato CSV. A coleta dos tuítes foi realizada pela equipe de pesquisa liderada pelas professoras Maria Camila Nardini Barioni e Elaine Ribeiro de Faria Paiva, docentes da Faculdade de Computação, na Universidade Federal de Uberlândia. Durante a coleta foram consideradas publicações, na Língua Portuguesa, que ocorreram durante a pandemia da COVID-19, tema foco da pesquisa da equipe mencionada. Para a extração em si, utilizou-se a biblioteca snscrape, disponível

na linguagem de programação Python. O código desenvolvido (Figura 5) para isso, considerou os seguintes parâmetros para a busca: extração em formato JSON, data de início da consulta, *twitter-search*: palavra-chave da consulta, until: data final da consulta, lang: idioma dos tuítes da consulta, near: localização da consulta e palavra-chave: COVID.

Figura 5 – Exemplo de código em Python para busca de tuítes.

```
os.system('snscrape --jsonl
                   --since: 2020-02-01
                   twitter-search: covid
                   until:2020-03-31
                   lang:pt
                   near:Brazil" > db.json')
```

Fonte: autoria própria

Para a realização do experimento, o presente trabalho utilizou duas bases diferentes, extraídas conforme o processo explicado anteriormente. A primeira foi utilizada na etapa de aperfeiçoamento do algoritmo de geocodificação, e a segunda, na aplicação de uma análise de sentimentos com nuvens de palavras. As características e períodos de extração de ambas as bases constam na Tabela 2.

Tabela 2 – Informações básicas das bases de tuítes utilizadas

Base	Quantidade de Tuítes	Período de Extração
1	4.977	27/04/2021 a 30/04/2021
2	261.341	02/2020 a 03/2020 06/2020 a 07/2020 11/2020 a 12/2020 01/2021 a 02/2021 06/2021 a 07/2021 12/2021 a 01/2022 04/2022 a 05/2022

Após isso, foram definidos quais campos contidos nas bases de dados seriam utilizados para o restante da pesquisa, que foram o ID, para poder associar a localização ao conteudo dos tuítes posteriormente, e a própria localização (parâmetro *Location* do campo *User*). Por meio da linguagem Python, os JSONs foram processados, o que possibilitou o acesso aos campos supracitados. Uma vez que o *Location* é de livre compartilhamento e edição do usuário, foram realizados vários testes com a base de tuítes adquirida para otimizar a acuracidade do algoritmo na rotulação do campo analisado.

4.2 Desenvolvimento de algoritmo de geocodificação com *Fuzzy Wuzzy*

Com os dados disponíveis, foi possível desenvolver e aprimorar um algoritmo de geocodificação do campo *Location* com a utilização da biblioteca *Fuzzy Wuzzy*, da linguagem Python, especialmente no processamento de tuítes brasileiros. Para isso, foram desenvolvidas duas funções principais que recebem o campo *Location* das bases importadas dos tuítes: a "fillVmatch" e a "fillVName" (Figura 6). A primeira recebe o campo *Location* e armazena na variável "word_to_check", então, a partir da função process.extract, da biblioteca *Fuzzy Wuzzy*, procura-se na lista "result", que é a lista de cidades brasileiras, se existir, a cidade com a maior similaridade considerando a faixa de similaridade testada. Após isso, apenas retorna a similaridade correspondente à cidade encontrada, ou retorna 0, quando não encontrada correspondência. Já a segunda função realiza o mesmo processo, retornando, no entanto, a cidade referente à similaridade retornada na primeira função, ou um texto vazio, nos casos sem correspondência.

Figura 6 – Funções que implementam a *Fuzzy Wuzzy* para criação do método de geocodificação.

```
def fillVmatch(x):
    word_to_check = x
    if not word_to_check:
        return 0
    a = process.extract(word_to_check,result,limit=1)
    if a[0][1]>= 86 and a[0][1]<= 100:
      # print('----FILLVMATCH----' + x + ' ' + str(a[0][1]))
        return a[0][1]
    else:
        return 0
def fillVName(c):
    word_to_check = c
    if not word_to_check:
      return 0
    a = process.extract(word_to_check,result, limit=1)
    if a[0][1]>= 86 and a[0][1]<= 100:
      # print('----FILLVNAME----' + c + ' ' + str(a[0][0]))
      return a[0][0]
    else:
        return ''
```

Fonte: autoria própria

As funções e processos descritos no último parágrafo foram colocados em um laço

para percorrer todos os tuítes das bases utilizadas. Ao final, foi possível gerar um *dataframe* com o resultado final de todos os tuítes, conforme fragmento de código da Figura 7, e dessa forma, exportá-lo para um arquivo em formato de planilha eletrônica, o que possibilitou a análise dos resultados.

Figura 7 – Dataframe sendo criado para receber resultado do processamento do campo *Location*.

```
df2 = pd.DataFrame(columns=('Campo Location', '%Similaridade', 'Local_Sugerido_Cidade','ID_correto'))
```

Fonte: autoria própria

4.3 Valor Preditivo Positivo (VPP)

Para avaliar o aperfeiçoamento contínuo do desenvolvimento e chegar a uma acuracidade satisfatória, calculou-se o Valor Preditivo Positivo, indicador também utilizado por Jiang et al. (2020), para apurar os resultados de seus experimentos com tuítes e a localização dos usuários da base utilizada por eles. Trata-se da proporção de rótulos que apresentaram resultado correto na sugestão de localização realizada pelo código desenvolvido com relação ao total de sugestões (*True Positives/(True Positives + False Positives)*) (PATINO; FERREIRA, 2017). Em cada nova versão do código, o Valor Preditivo Positivo foi utilizado para validar a acuracidade retornada após percorrer todos os campos *Location* analisados. Nos casos em que a base 1 (Tabela 2) foi utilizada, a contagem dos falsos positivos (FPs) e verdadeiros positivos (TPs) foi realizada manualmente em planilha eletrônica (Figura 8). Na coluna "TESTE" é feita uma validação se as colunas "Local_Sugerido_Cidade" e "Campo Location" contém *strings* iguais para agilizar o processo. Já na coluna "ACERTO?", é definido se o resultado é um verdadeiro positivo ou não.

Figura 8 – Fragmento da planilha com os campos de validação manual do resultado retornado pelo código.

Campo Location	%Similaridade	Local_Sugerido_Cidade	ID_correto	TESTE	ACERTO?
mare rio de janeiro	90	sumare	13874225267756974	FALSO	NAO
mare rio de janeiro	90	sumare	13874225199808675	FALSO	NAO
sumare brasil	100	sumare	13874184854883328	FALSO	SIM

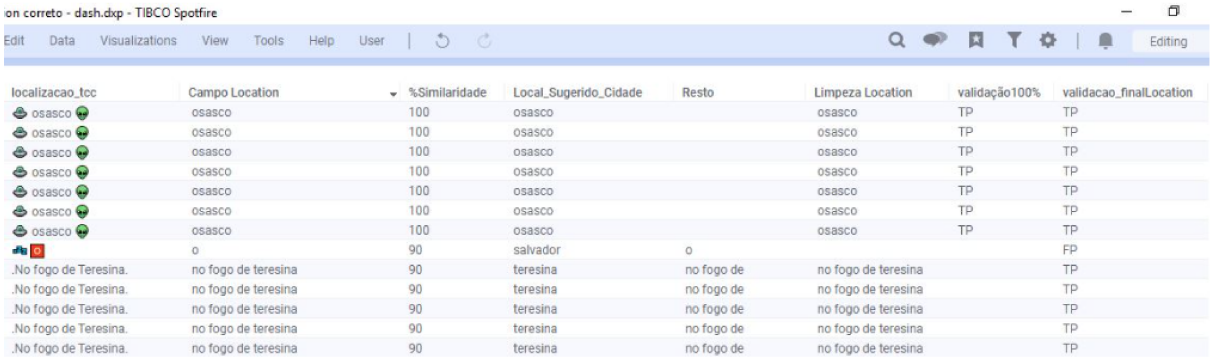
Fonte: autoria própria

Já no processamento da base 2 (Tabela 2), que é maior, foi utilizada a ferramenta *TIBCO Spotfire* para a mesma contagem, pois nela foi possível automatizar a validação das quantidades de TPs e FPs (GROUP, 2022)

A validação mencionada se consistiu nos seguintes passos: quando a similaridade obtida foi de 100%, não foi necessária nenhuma checagem manual, conforme exemplos não vazios da coluna "validação100%", mostrada na Figura 9. No restante, verificou-

se a diferença na *string*, ou seja, o conteúdo do campo *Location* que não era similar à cidade retornada, uma vez que essa diferença poderia ocasionar uma sugestão incorreta do *Fuzzy Wuzzy* ou, em outras palavras, falsos positivos. Um exemplo disso também aparece na Figura 9, quando o código sugeriu a cidade de Salvador, para uma *string* que continha apenas a letra "o". Sendo assim, a última coluna à direita, denominada "validacao_FinalLocation" indicou que se tratava de um falso positivo. Isso aconteceu porque ela foi gerada por meio da validação da presença ou não da *string* "salvador" no campo *Location* original, e no caso mostrado, não estava presente.

Figura 9 – Fragmento da ferramenta *TIBCO Spotfire* com os campos de validação.



localizacao_tcc	Campo Location	%Similaridade	Local_Sugerido_Cidade	Resto	Limpeza Location	validação100%	validacao_finalLocation
osasco	osasco	100	osasco		osasco	TP	TP
osasco	osasco	100	osasco		osasco	TP	TP
osasco	osasco	100	osasco		osasco	TP	TP
osasco	osasco	100	osasco		osasco	TP	TP
osasco	osasco	100	osasco		osasco	TP	TP
osasco	osasco	100	osasco		osasco	TP	TP
osasco	osasco	100	osasco		osasco	TP	TP
	o	90	salvador	o			FP
.No fogo de Teresina.	no fogo de teresina	90	teresina	no fogo de	no fogo de teresina		TP
.No fogo de Teresina.	no fogo de teresina	90	teresina	no fogo de	no fogo de teresina		TP
.No fogo de Teresina.	no fogo de teresina	90	teresina	no fogo de	no fogo de teresina		TP
.No fogo de Teresina.	no fogo de teresina	90	teresina	no fogo de	no fogo de teresina		TP
.No fogo de Teresina.	no fogo de teresina	90	teresina	no fogo de	no fogo de teresina		TP

Fonte: autoria própria

4.4 Teste de Similaridade

Com a finalidade de detectar a faixa ou valor especifico de similaridade de *strings* da biblioteca *Fuzzy Wuzzy* que retornaria o maior grau de assertividade de localidades sugeridas, considerou-se, inicialmente, as faixas de 65 a 75%, 76 a 85% e 86 a 100% de similaridade entre *strings*. Para isso, também foi adquirida uma base criada pelo IBGE (2022), com as 5.570 cidades brasileiras e suas populações estimadas, que o código desenvolvido utilizou para comparação entre *strings*. Conforme detalhado anteriormente, o código foi desenvolvido de forma a retornar o valor da similaridade em porcentagem, a localidade da base de cidades brasileiras referente à similaridade retornada, ou seja, a cidade sugerida pelo método, e o ID do tuíte para posterior validação. Caso não fosse encontrada nenhuma similaridade entre o texto contido no campo *Location* e alguma das cidades brasileiras, nas faixas de similaridades propostas, o código retornaria valor nulo nos campos descritos.

4.5 Conversão de letras maiúsculas para minúsculas

Esse processamento foi realizado após a definição da melhor faixa de similaridade. Tal teste se resumiu na transformação tanto das bases de pesquisa de localidade quanto

do campo *Location* para *strings* em letra minúscula. Com esse processamento, a intenção foi entender até que ponto isso diminuiria as diferenças entre as *strings* comparadas, e consequentemente, aumentaria a precisão da codificação realizada até aquele ponto.

4.6 Remoção de caracteres especiais e pontuação

Consistiu na remoção de acentuação, pontuação e caracteres especiais do campo *Location* e também das bases de comparação com cidades e estados. A finalidade deste teste foi a mesma do teste anteriormente apresentado.

4.7 Identificação de oportunidades de melhoria

Após realizar os testes de processamento anteriores e aplicá-los no algoritmo desenvolvido, a identificação de oportunidades de melhoria no processamento do campo *Location* foi realizada. Por meio da análise manual dos resultados, foi possível encontrar padrões de escritas e informações que confundiam o código desenvolvido, e dessa forma, criar funções para pre-processar as *strings* e, como consequência, fazer com que a quantidade de sugestões realizadas corretamente fosse maior.

Foi possível, por exemplo, tratar casos em que os usuários registraram as cidades, juntamente com as *strings* "Rio de Janeiro" e "São Paulo", sendo que a intenção deles foi registrar também o estado onde a cidade se situa. Nas primeiras versões do código, notou-se que, nesses casos, a sugestão retornava os nomes desses estados, por também serem as capitais dos mesmos, e não a cidade do usuário em si, o que gerava grande quantidade de falsos positivos. Pode-se citar, por exemplo, *strings* como "Volta Redonda Rio de Janeiro" ou "Barueri Sao Paulo". Esses problemas foram resolvidos com a criação e utilização de funções (Figura 10) capazes de checar a existência de *strings* como "rio de janeiro", "sao paulo", "brasil" e "minas gerais", no campo *Location*, com algum outro conteúdo, que poderia então, ser o nome de outra cidade. Nesses casos, as *strings* foram substituídas por uma *string* vazia para então serem processadas pela *Fuzzy Wuzzy*, gerando, portanto, o resultado correto.

Figura 10 – Fragmento do código com funções utilizadas para pré-processamento.

```
25 def substituirBR(x):
26     word_to_check = x
27     resultado = word_to_check.replace("brasil","")
28     encontrarBSB = word_to_check.find("brasilia")
29     if encontrarBSB >= 0:
30         return word_to_check
31     elif encontrarBSB < 0 and resultado == "" or resultado == " ":
32         return ""
33     else:
34         return resultado
35
36
37 def findAndReplace2(x):
38   word_to_check = x
39   res = word_to_check.find("sao paulo")
40   res2 = word_to_check.find("rio de janeiro")
41   res3 = word_to_check.find("brasil")
42   res4 = word_to_check.find("minas gerais")
43
44
45   if res < 0 and res2 >= 0 and res3 < 0 and res4 < 0:
46       return substituirRJ(word_to_check)
47   elif res >= 0 and res2 < 0 and res3 < 0 and res4 < 0:
48       return substituirSP(word_to_check)
49   elif res < 0 and res2 < 0 and res3 >= 0 and res4 < 0:
50       return substituirBR(word_to_check)
```

Fonte: autoria própria

Além da tratativa exemplificada acima, a função parcialmente mostrada na Figura 10, "findAndReplace2" também tratou situações em que não se encontrou nenhuma das *string* mencionadas e outras possibilidades, visando maior assertividade na sugestão retornada pelo código.

Na figura também é possível observar a função "substituirBR", utilizada pela "findAndReplace2", uma vez que percebeu-se também a sugestão incorreta de localidade quando havia "brasil" no conteúdo do campo *Location*, retornando resultados como "Americo Brasiliense", por exemplo. Ademais, considerou-se a possível presença do nome "brasilia", que poderia ser afetada pela função, já que poderia remover o fragmento "brasil", retornando apenas "lia" para ser processado. Sendo assim, considerou-se na função que o campo não poderia ser processado.

Enfim, a intenção dessa etapa foi diminuir ao máximo a quantidade de falsos positivos retornados por meio do tratamento dos casos mais comuns de dados de entrada, que conforme analisado nos resultados anteriores, foram capazes de confundir o processo de cálculo de similaridade de *strings* por parte do algoritmo.

4.8 Aplicação em bases de tuítes de trabalho correlato

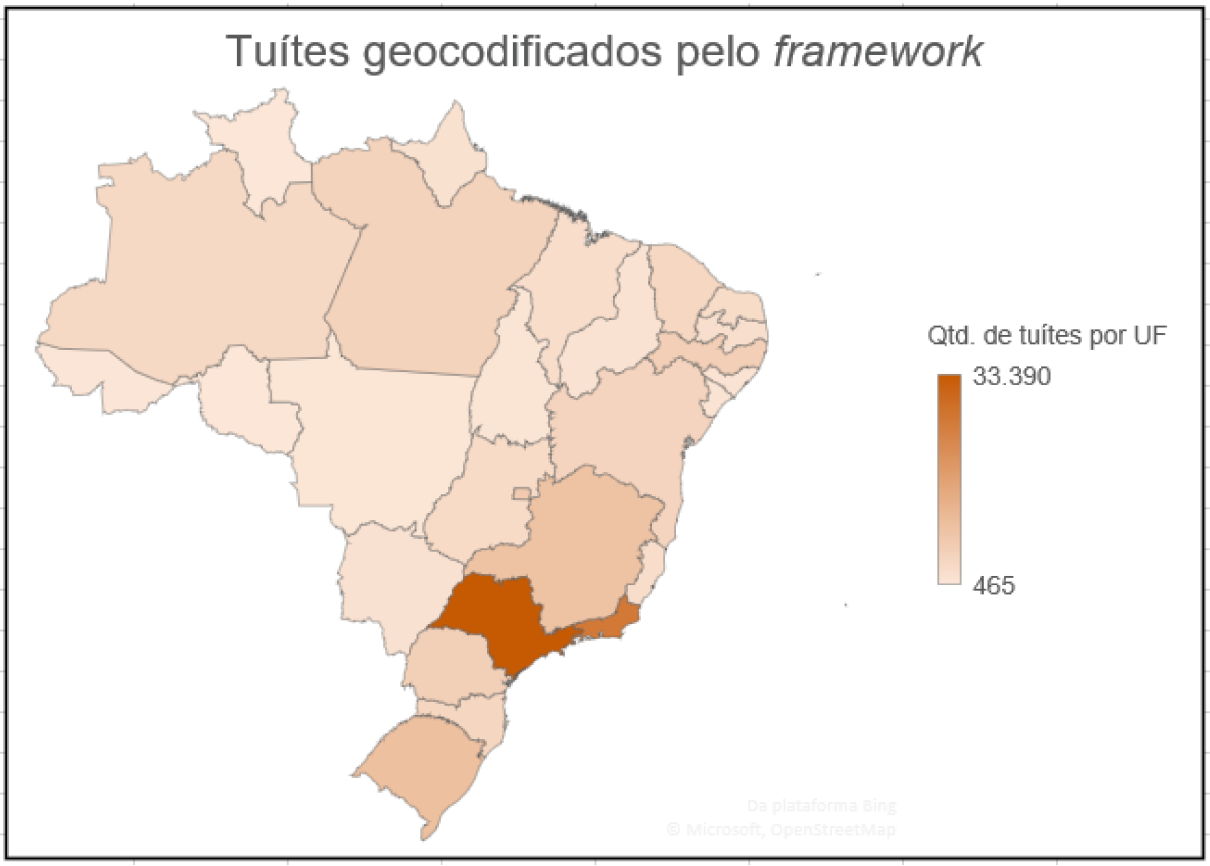
Com o código aperfeiçoado para tuítes do Brasil, e também, levando-se em conta tuítes em Língua Portuguesa, foi possível aplicá-lo na base com 261.341 tuítes (Tabela

2). Dessa maneira, tornou-se possível representar os dados dos tuítes de forma associada com a localização dos usuários no Brasil, e assim, caracterizar cada região do país quanto ao que se pensou e compartilhou em relação à pandemia da COVID-19.

4.8.1 Representação dos dados: geração de nuvens de palavras

A aplicação do algoritmo na base de dados supracitada possibilitou a geocodificação dos tuítes, e consequentemente, a atribuição de cada um a uma cidade, a um estado brasileiro e a uma região do país. Com essa informação disponível para os verdadeiros positivos, ou seja, apenas para os tuítes que o código (*framework*) analisou e retornou uma sugestão de município correta, foi possível separar a base em cada uma das regiões e gerar nuvens de palavras para cada uma. O mapa da Figura x traz a quantidade por estado dos tuítes geocodificados pelo *framework*, de forma a indicar a distribuição geográfica dos tuítes verdadeiros positivos provenientes da base 2 (Tabela 2). Tais nuvens de palavras foram geradas a partir do campo textual, ou seja, a mensagem que cada usuário compartilhou em cada tuíte. Também foi realizada uma análise temporal com nuvens por meio da separação dos tuítes regionalmente e anualmente. Para a região Sudeste, foi realizada uma análise temporal mais detalhada em pares de meses (Tabela 2), com tuítes do ano de 2020 até 2022, para analisar as mudanças de palavras-destaque ao longo do período. Por fim, gerou-se nuvens de palavras para as cinco cidades mais populosas de cada região, para verificar se o resultado seria muito diferente em relação às nuvens geradas para as regiões completas. O processo de geração das nuvens é detalhado a seguir.

Figura 11 – Distribuição geográfica dos tuítes verdadeiros positivos provenientes da geocodificação da base 2 (Tabela 2)



Fonte: autoria própria

4.8.2 Pré-processamento para a geração de nuvens de palavras

Após a separação das bases já mencionadas, foi realizado o pré-processamento dos tuítes, ou seja, do conteúdo textual compartilhado pelos usuários. As seguintes técnicas foram utilizadas para realizar o pré-processamento dos tuítes:

- Conversão dos emojis em *strings*: foi realizada a conversão dos emojis presentes nos tuítes em uma *string* com seu significado. Para realizar a conversão foi utilizada a biblioteca Emoji implementada em Python;
- Conversão em minúsculo: foi realizada a conversão de todas as letras que estavam em maiúsculas em minúsculas, o que padronizou as bases de dados;
- Eliminação das pontuações: foram eliminadas todas as pontuações presentes nos tuítes;
- Eliminação de usuários e urls: foram eliminados os usuários e links para outros sites;
- Eliminação de caracteres não-alfabéticos: foram eliminados todos os caracteres não existentes no alfabeto brasileiro;

- Eliminação das *stopwords*: foram eliminadas as palavras que são *stopwords*. A lista de *stopwords* foi encontrada em bibliotecas disponíveis para Python, além de terem sido acrescentadas outras *stopwords* manualmente, uma vez que se percebeu o aparecimento de palavras nas nuvens que não continham significado importante quando analisadas fora do contexto, como no caso de preposições, conjunções e verbos. Por exemplo: “assim”, “vamo”, “agora”, “tava” (KHANNA, 2021).

- Conversão das palavras em *tokens*: os tuítes foram convertidos em conjuntos de termos, cada termo recebe o nome de *token* e é separado por vírgula;

- Lematização dos *tokens*: os *tokens* foram reduzidos a sua forma base. Ou seja, foram retiradas todas as inflexões e substantivos dos *tokens*;

- Conversão de abreviações: foi realizada a conversão das abreviações em sua palavra normal, utilizando o auxílio de um dicionário próprio construído pelo autor.

4.8.3 Geração das nuvens de palavras

Para a criação das nuvens de palavras, utilizou-se a biblioteca *word_cloud*, implementada em Python, considerando uma quantidade de 40 palavras por imagem gerada. A biblioteca *word_cloud* é capaz de gerar nuvens de palavras a partir de determinado documento de entrada. Por isso, conforme já detalhado, foi necessário criar um arquivo para cada nuvem prevista, com todos os tuítes classificados. Na Figura 12, pode-se observar um fragmento do código que mostra a função utilizada para realizar a junção de todos os conteúdos dos tuítes em um documento textual.

Figura 12 – Exemplo de código para criação do documento com todos os tuítes.

```
texto = []
def creatDocument(text):
  text_temp = str(text)
  texto.append(text_temp)
  return texto
```

Fonte: autoria própria

Após essa etapa, no código, o documento criado foi passado como parâmetro na função que realmente gera a nuvem de palavras. Pode-se observar, na Figura 13, o fragmento da função utilizada para realizar a criação do documento, sendo que, document_create é o documento criado anteriormente, background_color definiu a cor de fundo da nuvem de palavras e max_words definiu a quantidade de palavras na nuvem de palavras.

Figura 13 – Fragmento do código com a criação da variável que passou a conter uma nuvem.

```
wordcloud = WordCloud(max_words=40, background_color='white').generate(str(document_create))
```

Fonte: autoria própria

Por fim, cada nuvem foi plotada utilizando-se a biblioteca *matplotlib.pyplot*, que funciona como o MATLAB, conforme fragmento de código da Figura 14.

Figura 14 – Código com configuração para plotagem das nuvens de palavras.

```
fig = plt.figure(1, figsize=(15, 15))
plt.axis('off')
plt.imshow(wordcloud)
plt.show()
```

Fonte: autoria própria

Esse processo foi utilizado para gerar todas as nuvens e abordagens mencionadas no tópico 3.8.1. Isso possibilitou a análise e comparação de resultados entre as nuvens, com o intuito de encontrar diferenças ou informações significativas sobre os dados manipulados ao longo deste trabalho.

5 Resultados e Discussão

Neste capítulo, primeiramente, são apresentados todos os resultados dos testes de similaridade realizados, bem como dos testes para aprimoramento do *framework* que consta na Figura 4. Logo após isso, são apresentadas todas as nuvens de palavras geradas após aplicação do *framework* na base de tuítes relacionada à pandemia da COVID-19 (base 2, Tabela 2). Dessa maneira, com os tuítes classificados por região via tal *framework* e nuvens disponíveis, foi possível discutir todos os resultados na seção de Discussões, onde é realizada também, a análise de sentimentos das nuvens de palavras por meio da análise interpretativa das mesmas.

5.1 Testes de Similaridade

Para documentar os resultados dessa etapa, foram elaboradas as Tabelas 3, 4, 5, 6 e 7, que também trazem as variáveis alteradas em cada versão do código, o que possibilitou a discussão sobre tais resultados. No primeiro teste (Teste 1, Tabela 3), realizou-se a geocodificação dos campos *Location* com a biblioteca *Fuzzy Wuzzy* sem qualquer pré-processamento. Já no Teste 2 (Tabela 4), fez-se o mesmo, mas com o campo *Location* e base de municípios em letra minúscula. Nos testes 3 e 4, (Tabelas 5 e 6, respectivamente), testou-se o impacto de remover ou não os caracteres especiais e acentos nas mesmas bases, ou seja, na base de tuítes por meio do campo *Location* e na base de pesquisa com mais de 5 mil municípios brasileiros. Por fim, na Tabela 7, Teste 5, constam os resultados da mesma base inicial sendo submetida ao *framework* completo, após diversas mudanças, conforme detalhado na seção 4.7, que consta no capítulo 4 (Método de Trabalho).

Para melhor entendimento, nas tabelas mencionadas aqui, a coluna “Faixa de Similaridade(%)” detalha a faixa de similaridade de *strings* considerada naquela linha de resultado, a coluna FP mostra a quantidade de registros falso positivos, ou seja, municípios erroneamente sugeridos pelo *framework* e a coluna TP mostra a quantidade de registros verdadeiros positivos, ou seja, municípios corretamente sugeridos pelo *framework*. Por sua vez, a coluna “Sem Sugestão” diz respeito à quantidade de tuítes que *framework* não receberam sugestão de município, tendo em vista o conteúdo do campo *Location*. Na coluna “Total de sugestões”, consta a quantidade de sugestões realizadas, sejam elas corretas ou não. Finalmente, a coluna “VPP”, mostra o resultado geral do teste, a partir do cálculo do indicador detalhado na seção 4.3 deste trabalho.

Tabela 3 – Resultados do primeiro teste

Teste 1: Dados sem pré-processamento Qtd. de registros testados: 2499					
Faixa de Similaridade (%)	FP	TP	Sem Sugestão	Total de sugestões	VPP (%)
65 a 75	0	0	2499	0	0
76 a 85	0	0	2499	0	0
86 a 95	938	657	904	1595	41.2

Tabela 4 – Resultados do segundo teste

Teste 2: Campo *Location* e base de municípios em letra minúscula Qtd. de registros testados: 2499					
Faixa de Similaridade (%)	FP	TP	Sem Sugestão	Total de sugestões	VPP (%)
65 a 75	118	0	2381	118	0
76 a 85	29	0	2470	29	0
86 a 100	1071	1262	166	2333	54.1

Tabela 5 – Resultados do terceiro teste

Teste 3: Remoção de acentuação e caracteres especiais apenas na base de tuítes Qtd. de registros testados: 2499					
Faixa de Similaridade (%)	FP	TP	Sem Sugestão	Total de sugestões	VPP (%)
65 a 75	126	0	2373	126	0
76 a 85	36	8	2455	44	18.2
86 a 100	1592	722	185	2314	31.2

Tabela 6 – Resultados do quarto teste

Teste 4: Remoção de acentuação e caracteres especiais na base de tuítes e na base de municípios Qtd. de registros testados: 2499					
Faixa de Similaridade (%)	FP	TP	Sem Sugestão	Total de sugestões	VPP (%)
65 a 75	190	0	2309	190	0
76 a 85	39	1	2459	40	2.4
86 a 100	669	1581	249	2250	70.3

Tabela 7 – Resultados do quinto teste

Teste 5: Diversas mudanças Qtd. de registros testados: 2500					
Faixa de Similaridade (%)	FP	TP	Sem Sugestão	Total de sugestões	VPP (%)
90 a 100	38	1266	1196	1304	97.1

5.2 Aplicação em bases de tuítes de trabalho correlato

Após a realização de melhorias no *framework* de geocodificação criado, a base 2 (Tabela 2) passou pelo mesmo processamento proposto, ou seja, tal base, que é referente aos tuítes relacionados à pandemia da COVID-19, foi submetida ao código final, representado no *framework* completo que consta na Figura 4. Dessa maneira, obteve-se um conjunto de tuítes georreferenciados em granularidade em nível de município, o que possibilitou também, sua associação às 5 regiões do território brasileiro e subsequente aplicação em análise de sentimentos. A Tabela 8, referente ao Teste 6, mostra o resultado disso.

Tabela 8 – Resultados do processamento da base 2

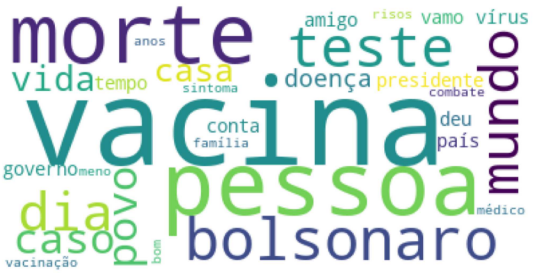
Teste 6: Processamento da base 2 Qtd. de registros testados: 261.341					
Faixa de Similaridade (%)	FP	TP	Sem Sugestão	Total de sugestões	VPP (%)
90 a 100	4.535	142.078	114.728	146.613	97.0

5.3 Geração de nuvens de palavras

As figuras que compõem esta seção foram geradas, conforme aquilo que foi detalhado na subseção 4.8.1 deste trabalho. Em resumo, trata-se das NPs criadas a partir da utilização do conteúdo textual dos tuítes classificados como TP (verdadeiros positivos), após aplicação do *framework* na base 2 (Tabela 2), ou seja, utilizando apenas os tuítes georreferenciados corretamente.

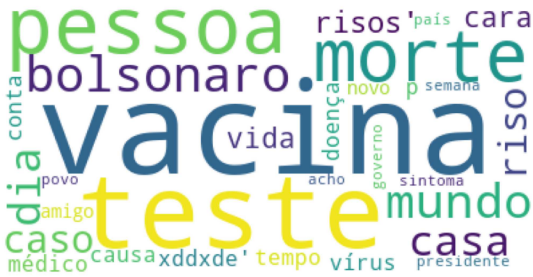
Figura 15 – Nuvens das regiões brasileiras em geral. **(a)** Centro Oeste **(b)** Nordeste **(c)** Norte **(d)** Sudeste **(e)** Sul

(a)

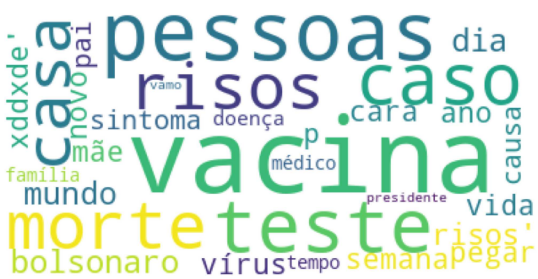


(b)

(c)

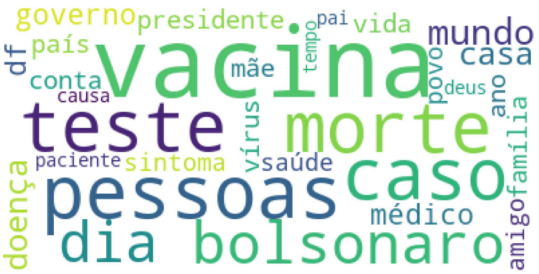


(d)

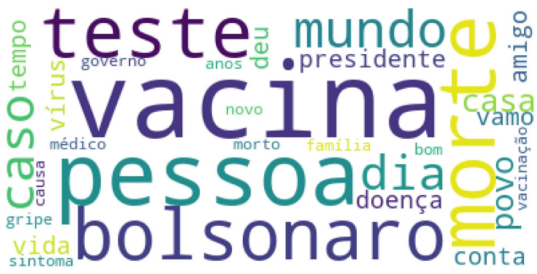


(e)

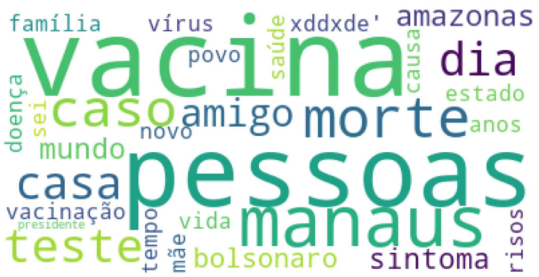
Figura 16 – Nuvens das regiões brasileiras por top 5 de cidades mais populosas. **(a)** Centro Oeste **(b)** Nordeste **(c)** Norte **(d)** Sudeste **(e)** Sul



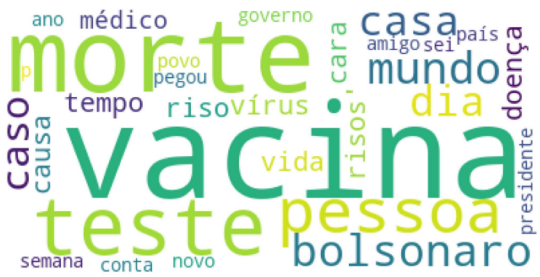
(a)



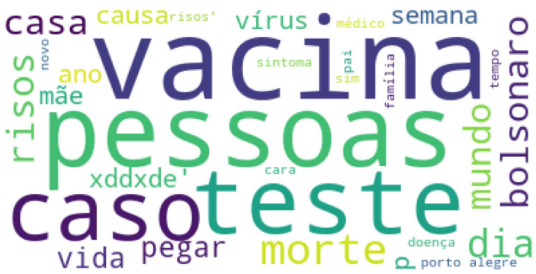
(b)



(c)



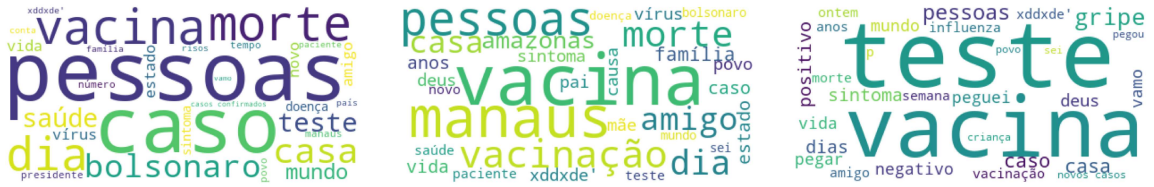
(d)



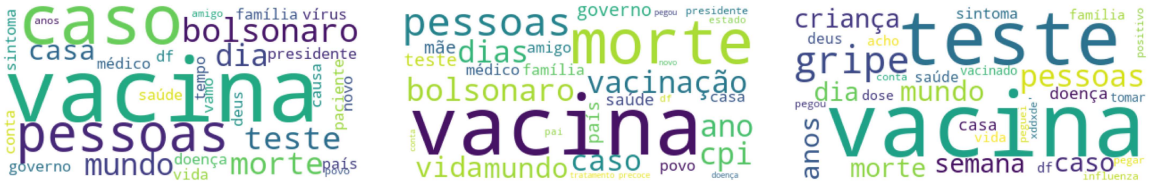
(e)

Figura 17 – Nuvens geradas por ano e região **(a)** Nuvens do Norte **(b)** Nuvens do Centro Oeste **(c)** Nuvens do Nordeste **(d)** Nuvens do Sudeste **(e)** Nuvens do Sul

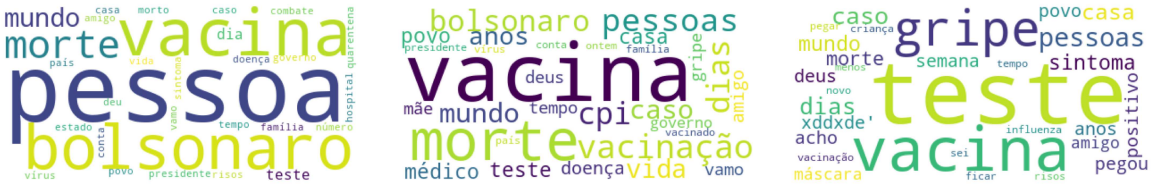
(2020) **(2021)** **(2022)**



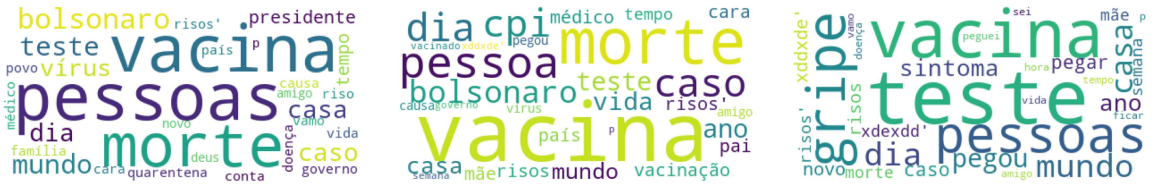
(a)



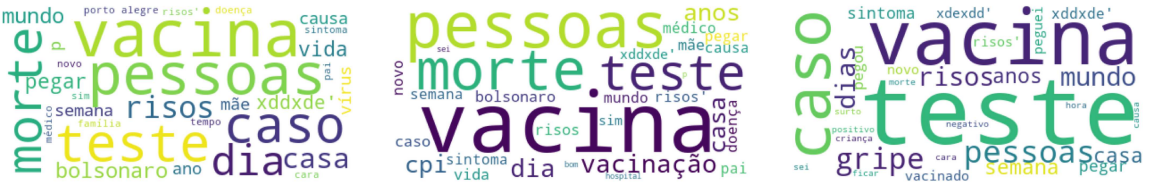
(b)



(c)

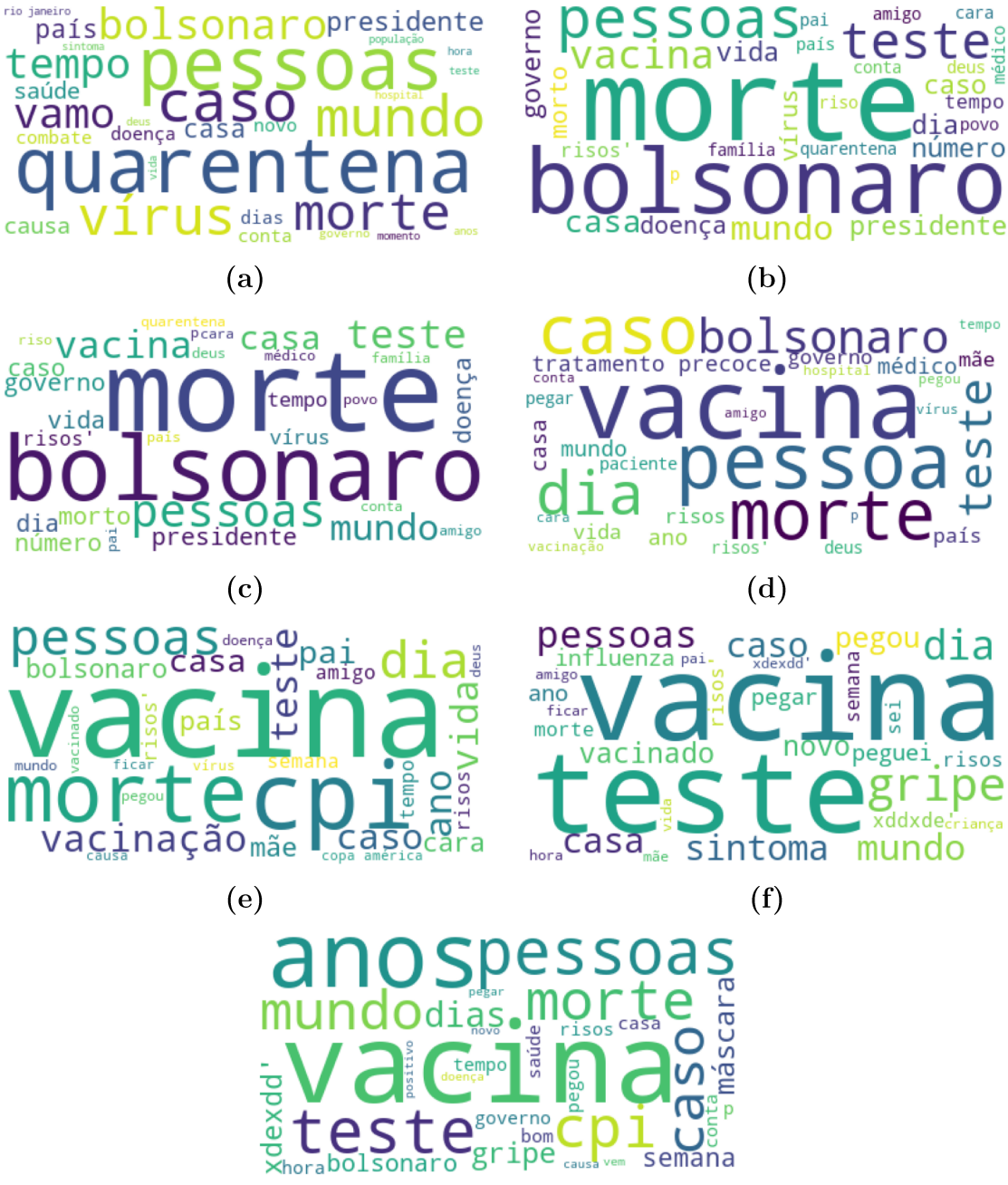


(d)



(e)

Figura 18 – Nuvens da região Sudeste por período de coleta. **(a)** Fev a Mar/2020 **(b)** Jun a Jul/2020 **(c)** Nov a Dez/2020 **(d)** Jan a Fev/2021 **(e)** Jun a Jul/2021 **(f)** Dez/2021 a Jan/2022 **(g)** Dez/2021 a Jan/2022



5.4 Discussões

Percebe-se, pela Tabela 3, que a precisão, ou VPP, do primeiro teste foi bem baixo (41,2%) ou zerado para as faixas de similaridade testadas. Esse resultado era esperado pelos seguintes motivos: falta de pré-processamento do campo *Location* e grande disponibilidade de cidades na base de municípios para comparação. Foi também nesse teste que

se percebeu a importância de considerar a faixa entre 96 e 100% de similaridade, já que com ela se espera o aumento das quantidades de verdadeiros positivos (TPs).

De acordo com o resultado do primeiro teste, foram realizadas duas alterações no segundo teste: descapitalização do campo *Location* e da base com mais de 5.000 municípios brasileiros e inclusão da faixa de similaridade de 96 a 100%. O resultado disso consta na Tabela 4. Dessa vez, para faixas menores de similaridade, o código conseguiu sugerir municípios, porém, devido à baixa similaridade, nenhuma sugestão foi um verdadeiro positivo. A faixa de 86 a 100%, porém, apresentou uma melhora de 12,9% no VPP com relação ao teste anterior, considerando a faixa de 86 a 95%, apresentando 54,1% de precisão.

A próxima alteração realizada no código foi a remoção de caracteres especiais e acentuação no campo *Location* da base de tuítes (Tabela 5) e, logo em seguida, fez-se também um teste removendo também essas características da base de municípios (Tabela 6). Segundo o resultado de tais tabelas, pode-se concluir que a remoção dos caracteres especiais e da acentuação teve um impacto considerável no VPP, uma vez que foi capaz de diminuir a assertividade do processamento do *Fuzzy Wuzzy* no terceiro teste com apenas o campo *Location* pré-processado. Além disso, com ambos os parâmetros, campo *Location* e base de municípios submetidos à remoção mencionada (Tabela 5), obteve-se um VPP de 70,3%, o que é uma melhoria de 16,2%, se comparado com o segundo teste (Tabela 4). O resultado obtido após os testes já discutidos corrobora com o conhecido problema de correspondência de *strings* da área de geotecnologias (JR; SALLES, 2007). Por saberem que a distância de Levensthein é uma abordagem afetada por acentuação e caracteres especiais, Jr e Salles (2007), precisaram adaptar tal método de correspondência de *strings* para nomes geográficos com um conjunto de caracteres especiais e acentos organizados em grupos, sendo que todos os componentes de um mesmo grupo foram considerados equivalentes. A adaptação foi realizada porque, segundo os pesquisadores, em algumas línguas, tais características podem ser determinantes para se encontrar uma correspondência. No caso do presente trabalho, mesmo que o *Fuzzy Wuzzy* use como base a distância de Levensthein, conforme explicitado na Tabela 6, obteve-se uma grande melhora no VPP após a desconsideração de caracteres especiais e acentos.

A partir do fim do teste anterior, foi necessário procurar nos resultados quais foram as principais causas de sugestões falsas positivas, uma vez que os demais tipos de pré-processamento de *strings* poderiam alterar muito o nome dos municípios. A tokenização, por exemplo, removeria preposições em *strings* como “Rio de Janeiro”. Lematização e *Stemming* poderiam gerar *strings* diferentes das originais. Isso causaria uma diminuição nas correspondências, conforme relatado por Jr e Salles (2007) .

Dentre as principais causas de falsos positivos encontrados, destacam-se:

1. Campos com *string* “Brasil” gerando sugestões como “Nova Brasilândia d’ Oeste”;
2. Cidades pequenas como “Horizonte” sendo sugeridas no lugar de “Belo Horizonte”;
3. Outras cidades dos estados do Rio de Janeiro e São Paulo deixando de serem sugeridas devido às capitais de mesmo nome constando juntas no campo *Location*;
4. Similaridade em 86% gerando apenas falsos positivos.

Tendo em vista tais problemas, foram criadas funções e realizadas alterações nos parâmetros de entrada do processamento, tratando-os conforme a numeração anterior:

1. Remoção das *strings* “brasil” ou “brazil” (transformada em “brasil” via código) do campo *Location*, exceto quando existem na *string* “brasilia” (capital do Distrito Federal);
2. Consideração das 200 cidades mais populosas, uma vez que elas concentram mais de 50% da população brasileira (IBGE, 2022). O restante reside nas outras 5.300 cidades, aproximadamente, e portanto, com chances menores de possuírem usuários tuitando;
3. Para o terceiro ponto, criou-se a função *findAndReplace2*, que ao achar as *strings* “rio de janeiro”, “sao paulo”, “brasil” e “minas gerais”, se comportou de acordo. Por exemplo, são tratados casos em que mais de uma dessas *strings* são encontradas, limpando-as da *string* geral, já que trata-se de uma ambiguidade. Para casos individuais, por exemplo, quando foi encontrado apenas “rio de janeiro”, verificou-se a existência ou não de algum fragmento da *string* que pudesse conter o nome de outra cidade, e em caso positivo, substituiu-se o fragmento “rio de janeiro” por vazio, o que permitiu o retorno apenas do fragmento não conhecido para a geocodificação;
4. Nesse caso, foram consideradas apenas similaridades da faixa entre 90 e 100% para garantir o maior número de verdadeiros positivos possível.

Enfim, após tais mudanças, realizou-se o processamento dos mesmos 2.500 registros, conforme Tabela 7. Nesse teste, foi gerado um VPP de 97,1%, ou seja, um valor próximo de 100. Na área da Medicina, espera-se que o valor de VPP de um teste clínico seja o mais próximo possível de 100 e, nesse caso, eles classificam tal teste como padrão gold (PARIKH et al., 2008). Para corroborar com a satisfação com relação ao resultado do teste da Tabela 7, pode-se citar Jiang et al. (2020), que ficaram confiantes com VPPs de 96,3 ou 98,2% nos testes que realizaram com amostras ao rotularem tuítes em nível estadual. O fato de que se obteve um maior número de tuítes não rotulados neste teste, devido à desconsideração de mais de 5.000 municípios da base de comparação, não é tão

importante em um contexto estatístico, uma vez que, com os tuítes, sempre trabalhar-se-á com uma amostra representativa da população, e não com a população-objeto, ou seja, toda a população brasileira (NETO, 2002).

Com os resultados obtidos até aqui, construiu-se o *framework* básico para geocodificação de tuítes brasileiros com o uso do campo *location* (Figura 4).

5.4.1 Aplicação em base de tuítes de trabalho correlato

Conforme definição do *framework*, foi possível realizar o processamento da base 2 (Tabela 2) que contém um número maior de registros. Apesar dessa característica, com a validação via *Spotfire* (GROUP, 2022), obteve-se um VPP de 97%, muito próximo daquele encontrado para a base de teste (base 1). Com isso, foi possível utilizar a base com verdadeiros positivos na geração de nuvens de palavras, o que permitiu então a realização da análise interpretativa subsequente.

5.4.2 Análise Interpretativa

Por meio das NPs contidas na Figura 14, pode-se notar interesses e comportamentos em comum entre os estados brasileiros, bem como diferenças. Por apresentarem um tamanho maior, as palavras “vacina” e “pessoas” foram os maiores tópicos de interesses dos brasileiros que comentaram sobre a Covid-19, em todas as regiões. Com esse resultado, é possível considerar que a demanda por vacinas e a preocupação com as pessoas foram assuntos muito debatidos no Twitter.

As diferenças também são perceptíveis quando se analisa a NP da região Norte (Figura 15 (c)), por exemplo, que tem como destaque as palavras “Manaus” e “Morte”. Foi em Manaus que ocorreu o epicentro da pandemia de Covid em 2021, o que de acordo com Silva et al. (2022), expôs a gravidade da pandemia em contextos de grande desigualdade social e fraca efetividade de ações governamentais. Dessa maneira, a NP foi capaz de destacar a preocupação com a situação em Manaus.

Outra diferença está no distinto destaque do sobrenome do Presidente da República na época, “Bolsonaro”, entre regiões, o que indica uma polarização política. Quando se compara as NPs das regiões Sul e Nordeste, no Sul o sobrenome está escrito em tamanho menor. Tal resultado vai de encontro com o resultado das eleições de 2022 no Brasil, em que o presidente da época, Jair Messias Bolsonaro, apresentou maior apoio na região Sul e maior desaprovação na região Nordeste, de acordo com a quantidade de votos (G1, 2022). Além disso, por se tratar de uma análise com tuítes relacionados à Covid-19, pode-se inferir que o ex-presidente tenha sido mais culpabilizado pela situação da pandemia no Brasil por parte dos nordestinos, em comparação com os habitantes do Sul.

Em uma análise comparativa com as Figuras 15 e 16, sendo a segunda referente às

5 cidades mais populosas de cada região, nota-se uma grande semelhança entre as palavras maiores, ou seja, mais frequentes. Isso mostra que a amostragem com menores localidades pode ser utilizada como forma de representar cada região.

Para uma análise temporal, gerou-se as NPs da Figura 17. Nesse caso, é possível indicar uma mudança de demandas e preocupações com o avanço da pandemia entre regiões e dentro de uma mesma região. Verifica-se de 2020 para 2021, em todas as regiões, um maior destaque da palavra "vacina", que provavelmente virou tendência devido ao início da vacinação no fim de 2020, algo que ocorreu no Reino Unido, em 8 de Dezembro de 2020 (CNN, 2020). No mesmo ritmo da vacinação, e consequente atenuação da pandemia, percebe-se o aparecimento da sigla "CPI" nas NPs de 2021, ano exato em que se ocorreu a CPI da Covid, criada para apurar a atuação do Governo Federal durante o evento epidemiológico (UOL, 2022). Ainda na Figura 17, em todas as regiões, percebe-se uma mudança com o destaque da palavra "teste" no ano de 2022, se comparado com os anos anteriores, o que possibilita diversas interpretações. Não é possível associar essa mudança a uma demanda ou razão específica, já que a testagem contra a Covid foi importante durante toda a pandemia para a identificação de transmissores e o isolamento oportuno dos mesmos (MEDICINASA, 2021).

Por fim, por meio da Figura 17 ainda foi possível validar a eficácia do *framework* criado na presente pesquisa. Isso porque nota-se o destaque de certas localizações geográficas nas NPs das regiões onde são encontradas. Por exemplo, "Manaus" aparece na nuvem da região Norte (Figura 17, (a), 2021). Já "Porto Alegre" é uma das palavras presentes na nuvem da região Sul (Figura 17, (e), 2020) e "DF", sigla para Distrito Federal, aparece nas nuvens da Região Centro-Oeste (Figura 17, (b)). Em outras palavras, o conteúdo compartilhado no campo textual pelos usuários consegue se relacionar com as regiões que eles compartilham no campo *Location*, o que traz certo nível de confiança no uso do *framework* proposto.

Finalmente, com relação à Figura 18, tem-se uma análise temporal diferente. Trata-se de NPs de recorte temporal para a região Sudeste. Nessas NPs é possível observar a evolução dos tópicos mais discutidos durante a pandemia na região Sudeste. "Quarentena", por exemplo, foi uma palavra em destaque na primeira nuvem, considerando o tempo de forma cronológica, e isso denota uma reação da população com a necessidade de se iniciar a quarentena no início do evento epidemiológico (Figura 18, (a)). Com o avanço da pandemia, a palavra "morte" passa a ter mais destaque, possivelmente devido ao aumento do número de mortes relacionadas à Covid-19. "Vacina" é a expressão que mais se destaca a partir de 2021, devido ao motivo já discutido anteriormente.

Todas as constatações feitas aqui reforçam a capacidade do *framework* criado no presente trabalho em associar a localização geográfica com os assuntos mais compartilhados no Twitter, bem como demonstra que, em associação com técnicas de análise de

sentimentos, pode-se obter *insights* interessantes sobre as opiniões, demandas ou anseios da população brasileira.

6 Conclusões e Recomendações para Trabalhos Futuros

Tendo em vista os objetivos deste trabalho, que foram propor um *framework* para detecção da localização geográfica de usuários a partir de metadados do Twitter no Brasil e ainda, avaliar a acuracidade daquilo que foi proposto, conclui-se que foram obtidos resultados satisfatórios. Isso porque obteve-se um VPP de 97,1% no teste final para detecção de localizações com granularidade de municípios, chegando portanto, próximo de um resultado considerado como padrão *gold* em outras áreas de pesquisa. Obteve-se também um VPP de 97%, quando se aplicou o *framework* em uma base de dados de trabalho correlato. Além de atingir os objetivos propostos, o trabalho alinhou-se com a justificativa apresentada ao utilizar o parâmetro *Location* do campo *User*, que é de livre preenchimento do usuário, como geolocalizador.

Ademais, os resultados obtidos permitiram a realização de uma análise de sentimentos em nível regional e temporal, com a geração de nuvens de palavras referentes aos tuítes geocodificados da base de dados do trabalho correlato. A análise interpretativa das nuvens geradas indicou que o conteúdo compartilhado no campo textual pelos usuários consegue se relacionar com as regiões que eles compartilham no campo *Location*, o que trouxe certo nível de confiança no uso do *framework* proposto. As constatações realizadas durante essa etapa reforçaram a capacidade do *framework* criado no presente trabalho em associar a localização geográfica com os assuntos mais compartilhados no Twitter, bem como demonstrou que, em conjunto com técnicas de análise de sentimentos, pode-se obter informações interessantes sobre a reação da população brasileira diante de eventos importantes, como uma pandemia, por exemplo.

Para esforços futuros, recomenda-se a utilização do *framework* proposto em múltiplas bases de dados do Twitter para se verificar qual é a média de precisão do *framework* quando submetido a dados de entrada diversificados, e dessa forma, encontrar maneiras de se melhorar a precisão de forma generalizada.

Recomenda-se ainda, a criação de *frameworks* que venham a empregar outras técnicas de correspondência de *strings* para a geocodificação de tuítes. Dessa forma, permitir-se-á a comparação com o que foi proposto no presente trabalho, com o intuito de se elencar a melhor maneira de executar a geocodificação.

Referências

ALESSA, A.; FAEZIPOUR, M.; ALHASSAN, Z. Text classification of flu-related tweets using fasttext with sentiment and keyword features. In: *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. [S.l.: s.n.], 2018. p. 366–367. Citado na página 21.

ASSIS, C. R. O. de. Análise comparativa de técnicas de classificação sobre uma ferramenta de detecção de empacotamento. *TCC - Sistemas de Informação (Monte Carmelo)*, Dezembro 2018. Citado na página 14.

BANNI, M. et al. Uma análise interpretativa pré- e intra-pandemia dos dados de redes sociais no domínio religioso. In: *Anais do XI Workshop sobre Aspectos da Interação Humano-Computador para a Web Social*. Porto Alegre, RS, Brasil: SBC, 2020. p. 1–8. ISSN 2596-0296. Disponível em: <https://sol.sbc.org.br/index.php/waihcws/article/view/12341>. Citado na página 21.

CNN. *Veja quais países iniciaram a vacinação contra a Covid-19; Brasil está fora*. 2020. <https://www.cnnbrasil.com.br/saude/quais-os-paises-que-ja-comecaram-a-vacinacao-contra-a-covid-19/>. (Acessado em 03/01/2023). Citado na página 43.

DIXON, S. J. *Number of Twitter users worldwide from 2019 to 2024*. 2022. (Acessado em 03/01/2023). Disponível em: <https://www.statista.com/statistics/303681/twitter-users-worldwide/#statisticContainer>. Citado na página 8.

FOUNDATION, P. S. *FuzzyWuzzy*. 2022. (Acessado em 03/01/2023). Disponível em: <https://pypi.org/project/fuzzywuzzy/>. Citado na página 16.

G1. *Eleito presidente, Lula venceu Bolsonaro no Nordeste; veja análise por região | Eleição em Números | G1*. 2022. <https://g1.globo.com/politica/eleicoes/2022/eleicao-em-numeros/noticia/2022/10/31/eleito-presidente-lula-so-venceu-bolsonaro-no-nordeste-veja-analise-por-regiao.ghtml>. (Acessado em 03/01/2023). Citado na página 42.

GOMAA, W.; FAHMY, A. A survey of text similarity approaches. *international journal of Computer Applications*, v. 68, 04 2013. Citado 2 vezes nas páginas 15 e 16.

GROUP, C. S. *Spotfire*. 2022. (Acessado em 03/01/2023). Disponível em: <https://www.tibco.com/pt-br/products/tibco-spotfire>. Citado 2 vezes nas páginas 25 e 42.

HAN, J.; KAMBER, M. *Data mining: Concepts and techniques*. Burlington, MA: Elsevier, 2011. ISBN 9780123814791 0123814790. Disponível em: <http://www.worldcat.org/search?qt=worldcat_org_all&q=9780123814791>. Citado na página 14.

HOU, Q.; HAN, M.; CAI, Z. Survey on data analysis in social media: A practical application aspect. *Big Data Mining and Analytics*, v. 3, n. 4, p. 259–279, 2020. Citado 5 vezes nas páginas 8, 11, 14, 18 e 21.

IBGE. *Estimativas Da População Residente No Brasil E Unidades Da Federação*. 2022. (Acessado em 03/01/2023). Disponível em: <https://ftp.ibge.gov.br/Estimativas_de_Populacao/Estimativas_2020/POP2020_20220711.xls>. Citado 2 vezes nas páginas 26 e 41.

INFOPEDIA. *Framework*. 2003–2021. (Acessado em 03/01/2023). Disponível em: <https://www.infopedia.pt/dicionarios/ingles-portugues/framework>. Citado na página 18.

JIANG, J. et al. Political polarization drives online conversations about covid-19 in the united states. *Human Behavior and Emerging Technologies*, v. 2, n. 3, p. 200–211, 2020. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/hbe2.202>. Citado 5 vezes nas páginas 8, 20, 21, 25 e 41.

JR, C. D.; SALLES, E. Approximate string matching for geographic names and personal names. In: . [S.l.: s.n.], 2007. p. 49–60. Citado na página 40.

KARAMI, A. et al. Twitter and research: A systematic literature review through text mining. *IEEE Access*, v. 8, p. 67698–67717, 2020. Citado na página 8.

KHANNA, C. *Text pre-processing: Stop words removal using different libraries*. 2021. (Acessado em 03/01/2023). Disponível em: <https://towardsdatascience.com/text-pre-processing-stop-words-removal-using-different-libraries-f20bac19929a>. Citado na página 31.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. (IJCAI'95), p. 11371143. ISBN 1558603638. Citado na página 14.

LIU, Z. et al. Analysis of the performance and robustness of methods to detect base locations of individuals with geo-tagged social media data. *International Journal of Geographical Information Science*, Taylor Francis, v. 35, n. 3, p. 609–627, 2021. Disponível em: <https://doi.org/10.1080/13658816.2020.1847288>. Citado na página 9.

MAJUMDER, P. *FuzzyWuzzy Python Library: Interesting Tool for NLP and Text Analytics*. 2021. (Acessado em 03/01/2023). Disponível em: <https://www.analyticsvidhya.com/blog/2021/06/fuzzywuzzy-python-library-interesting-tool-for-nlp-and-text-analytics/>. Citado na página 17.

MARTINS, J. S. et al. *Processamentos de Linguagem Natural*. [S.l.]: Soluções Educacionais Integradas, 2020. Citado 6 vezes nas páginas 11, 12, 13, 14, 15 e 16.

MEDICINASA. *Covid-19: Mesmo com a vacina, ainda é importante a testagem - Medicina S/A*. 2021. <https://medicinasa.com.br/importancia-testagem/>. (Acessado em 03/01/2023). Citado na página 43.

MELO, T. D. *Dataset of Tweets and News Media on COVID-19 in Portuguese*. Mendeley, 2020. (Acessado em 03/01/2023). Disponível em: <https://data.mendeley.com/datasets/vhxdgjfjnk/3>. Citado na página 9.

MOREIRA, S. F.; BAKLIZKY, M.; DIGIAMPIETRI, L. A. Uso de mineração de textos para a identificação de postagens com informações de localização. In: *Anais do VII Brazilian Workshop on Social Network Analysis and Mining*. Porto Alegre, RS, Brasil: SBC, 2018. ISSN 2595-6094. Disponível em: <https://sol.sbc.org.br/index.php/brasnam/article/view/3600>. Citado na página 20.

MUKHERJEE, I.; SAHANA, S.; MAHANTI, P. An improved information retrieval approach to short text classification. *International Journal of Information Engineering and Electronic Business*, v. 9, p. 31–37, 07 2017. Citado na página 14.

NETO, P. L. d. O. C. *Estatística*. São Paulo: Blucher, 2002. Citado na página 42.

NOW, D. T. *Hydrator [Computer Software]*. 2022. (Acessado em 03/01/2023). Disponível em: <https://github.com/docnow/hydrator>. Citado na página 9.

PARIKH, R. et al. Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, v. 56, n. 1, p. 45–50, 2008. Disponível em: <https://www.ijo.in/article.asp?issn=0301-4738;year=2008;volume=56;issue=1;spage=45;epage=50;aulast=Parikh;t=6>. Citado na página 41.

PATINO, C. M.; FERREIRA, J. C. Understanding diagnostic tests. part 2. *Jornal Brasileiro de Pneumologia*, FapUNIFESP (SciELO), v. 43, n. 6, p. 408–408, dez. 2017. Disponível em: <https://doi.org/10.1590/s1806-37562017000000424>. Citado na página 25.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 14.

REIS, J. et al. Breaking the news: First impressions matter on online news. In: . [S.l.: s.n.], 2015. Citado na página 18.

SALDANA-PEREZ, M. et al. When twitter becomes a data source for geospatial analysis. *Research in Computing Science*, v. 148, p. 357–374, 12 2019. Citado na página 20.

SILVA, L. E. P. D. et al. Amazonas no epicentro da pandemia de COVID-19 uma revisao sistemática / amazon at the epicenter of the COVID-19 pandemic a systematic review. *Brazilian Journal of Health Review*, South Florida Publishing LLC, v. 5, n. 3, p. 9270–9280, maio 2022. Disponível em: <https://doi.org/10.34119/bjhrv5n3-105>. Citado na página 42.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introdução ao datamining: mineração de dados*. Ciência Moderna, 2009. Hardcover. ISBN 0321321367. Disponível em: <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0321321367>. Citado na página 14.

TARDELLI, A.; DIAS, A.; FRANÇA, J. Introdução à análise de sentimentos com word clouds. In: _____. [S.l.: s.n.], 2019. p. 38–67. ISBN 9788576694885. Citado 2 vezes nas páginas 18 e 19.

TIBURCIO, G. V. Avaliação experimental de classificadores para análise de sentimentos em dados de redes sociais. *Trabalho de Conclusão de Curso (Graduação em Estatística)*, Outubro 2021. Citado 2 vezes nas páginas 13 e 14.

TUYCHIEV, B. *FuzzyWuzzy: Fuzzy String Matching in Python, Beginner's Guide*. 2020. (Acessado em 03/01/2023). Disponível em: <https://towardsdatascience.com/fuzzywuzzy-fuzzy-string-matching-in-python-beginners-guide-9adc0edf4b35>. Citado 2 vezes nas páginas 17 e 18.

UFABC. *CPI da Covid*. 2021. (Acessado em 03/01/2023). Disponível em: <https://observa.pesquisa.ufabc.edu.br/list/cpi-da-covid/>. Citado na página 9.

UOL. *O que aconteceu com as conclusões da CPI da Covid-19?* 2022. <https://noticias.uol.com.br/politica/ultimas-noticias/2022/10/29/o-que-aconteceu-com-as-conclusoes-da-cpi-da-covid-19.htm>. (Acessado em 03/01/2023). Citado na página 43.

ZHANG, W.; GELERNTER, J. Geocoding location expressions in twitter messages: A preference learning method. *Journal of Spatial Information Science*, v. 9, 12 2014. Citado 2 vezes nas páginas 12 e 15.

ZRIGUI, M. et al. Arabic text classification framework based on latent dirichlet allocation. *J. Comput. Inf. Technol.*, v. 20, n. 2, p. 125–140, 2012. Disponível em: <http://cit.srce.unizg.hr/index.php/CIT/article/view/1770>. Citado na página 21.