UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

ARTHUR FACIN DE BORTOLI

MODELO PARA EXTRAÇÃO DE DADOS E ANÁLISE DE SENTIMENTOS EM CIDADES INTERMEDIÁRIAS: UMA ABORDAGEM UTILIZANDO USUÁRIOS CENTRAIS DO *TWITTER*

DISSERTAÇÃO

PATO BRANCO

2023

ARTHUR FACIN DE BORTOLI

MODELO PARA EXTRAÇÃO DE DADOS E ANÁLISE DE SENTIMENTOS EM CIDADES INTERMEDIÁRIAS: UMA ABORDAGEM UTILIZANDO USUÁRIOS CENTRAIS DO *TWITTER*

Model for Data Extraction and Sentiment Analysis in Intermediate Cities: an approach using Twitter Central Users

Dissertação apresentada como requisito para obtenção do título de Mestre em Engenharia de Produção e Sistemas da Universidade Tecnológica Federal do Paraná (UTFPR).
Orientador(a): Gilson Ditzel Santos.

PATO BRANCO

2023

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Pato Branco

ARTHUR FACIN DE BORTOLI

MODELO PARA EXTRAÇÃO DE DADOS E ANÁLISE DE SENTIMENTOS EM CIDADES INTERMEDIÁRIAS: UMA ABORDAGEM UTILIZANDO USUÁRIOS CENTRAIS DO TWITTER

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Engenharia De Produção E Sistemas da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Gestão Dos Sistemas Produtivos.

Data de aprovação: 27 de Fevereiro de 2023

Dr. Gilson Ditzel Santos, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Gilson Adamczuk Oliveira, Doutorado - Universidade Tecnológica Federal do Paraná

Dra. Marie Anne Macadar Moron, Doutorado - Universidade Federal do Rio de Janeiro (Ufrj)

AGRADECIMENTOS

Agradeço aos colegas do PPGEPS, pela rica troca de ideias e experiências que tivemos ao longo destes dois anos.

Agradeço aos professores do programa, pelo empenho dispendido nas disciplinas e à secretária do Programa, que sempre me atendeu prontamente.

Agradeço à Banca que topou avaliar o trabalho, dando sugestões pertinentes, indicando correções e melhorias a serem feitas.

Agradeço ao meu Orientador, pelos quase sete anos de parceria, sendo o principal mentor da minha jornada acadêmica.

Aos colegas de trabalho da DIRPLAD, especialmente aos colegas de DEMAP, por terem aguentado minhas lamúrias com os estudos, durante o expediente, ao longo deste período.

Agradeço à UTFPR, minha segunda casa, desde 2015, como aluno, e desde 2019, como servidor, lugar de algumas poucas angústias, mas de grandes alegrias e realizações.

Agradeço à minha família, pelo suporte e compreensão de minhas participações exíguas na hora do mate. À minha irmã, a quem tento ser exemplo, à minha mãe, que ouve e compreende meus desabafos, e ao meu pai, esteio nas minhas decisões, de quem sigo o conselho “estude, menino! ”, desde sempre. Agradeço, também, à minha namorada, que, por compartilhar dos mesmos anseios e sonhos que eu, entendeu os vários momentos em que os afazeres do mestrado estiveram em primeiro plano, e razão pela qual, por vezes, estive com o pensamento distante.

Finalmente, e mais importante, obrigado, meu Deus! Obrigado por aliviar o fardo, quando recorri a Ti, por conduzir meus caminhos e iluminar minhas decisões, e por me conceder ânimo e força de vontade maiores do que desculpas para não seguir em frente.

RESUMO

Acredita-se que a população urbana global, no ano de 2050, alcance a marca de 68% do total de habitantes. Na esteira da crescente, rápida e contínua urbanização, tem-se a complexificação dos desafios de gestão, que tem feito com que as cidades, pensando em prover melhor qualidade de vida aos seus cidadãos, passem a adotar conceitos de Cidades Inteligentes. Uma cidade passa a ser razoavelmente tecnológica, sob a ótica de Cidades Inteligentes, quando é capaz de fazer uso das ferramentas de TIC para resolução de problemas urbanos. Enquanto subconjunto destas ferramentas de TIC tem-se as redes sociais, que são produtoras de dados contínuos, de ampla difusão e disseminação. A capacidade de conhecer sobre que assuntos os cidadãos mais estão discutindo, e como se sentem diante disso, por meio dos dados de redes sociais, é algo que vai ao encontro da utilização da TIC, voltada ao contexto de Cidades Inteligentes. Todavia, a tarefa de extração e análise de dados de redes sociais não é algo trivial, especialmente em cidades de porte intermediário, brasileiras, o que clama por novas soluções. Assim, o que se propõe aqui é um modelo para extração e análise de dados, tendo por fonte a rede social *Twitter*, geograficamente situados no contexto das cidades intermediárias, tendo a cidade de Maringá, como *lócus* do estudo. Para tanto, faz-se uso de uma abordagem em usuários centrais da rede social *Twitter*, para extração dos dados. Para análise, são aplicadas a Alocação Latente de Dirichlet (LDA), para identificar o(s) tópico(s) proeminente(s), e Análise de Sentimentos, para descobrir sua polaridade. A extração situada no contexto das cidades, fazendo uso tanto da abordagem de usuários centrais, como da abordagem geolocalizada foi bem-sucedida e a aplicação da LDA obteve êxito, indicando os tópicos mais discutidos, o que possibilitou a constatação das temáticas mais discutidas pelos *netizens.* Observou-se que os tópicos principais identificados por meio da LDA foram relacionados a: Pandemia, no primeiro teste de aplicação do modelo, com coleta em janeiro de 2022, e Mobilidade Urbana, com coleta feita de setembro a dezembro de 2022, no segundo teste de aplicação do modelo. Realizando a coleta temática acerca destes assuntos, a polaridade de sentimento predominante dos *netizens* sobre o tema foi negativa. Verificou-se ainda que a estratégia supervisionada foi mais bem sucedida para classificação da polaridade de sentimento dos *tweets*, sobretudo os métodos SVM e *Random Forest*, sobretudo no segundo teste de aplicação do modelo. Tendo o aporte da literatura e posterior a realização de dois testes, o fluxograma final bem como um modelo genérico são apresentados ao final, sendo, posteriormente, levantadas limitações do trabalho e oportunidades para estudos futuros.

Palavras-chave: Extração de dados; Cidades Inteligentes; Cidades Intermediárias; Modelagem de Tópicos; Análise de Sentimentos; Machine Learning; Twitter

ABSTRACT

It is believed that the global urban population, in the year 2050, will reach the mark of 68%. In the wake of the growing, rapid and continuous urbanization, management challenges have become more complex, which has made cities, thinking about providing a better quality of life for their citizens, start to adopt concepts of Smart Cities. A city becomes reasonably technological, from the point of view of Smart Cities, when it is able to make use of ICT tools to solve urban problems. As a subset of these ICT tools, there are social media, in which are produced continuous data, of wide diffusion and dissemination. The ability to know what subject citizens are talking about the most, and how they feel about it, through data from social media, is something that meets the use of ICT, aimed at the context of Smart Cities. However, the task of extracting and analyzing data from social media is not trivial, especially in Brazilian cities of intermediate size, which calls for new solutions. Thus, what is proposed here is a model for extracting and analyzing data, having the social media Twitter as the source of content, geographically located in the context of intermediate cities, with the city of Maringá as the locus of this study. For that, an approach of central users is used, in the social network Twitter, to extract the data. For analysis, Dirichlet Latent Allocation (LDA) is applied to identify the prominent topic(s), and Sentiment Analysis to discover its polarity. The data extraction located in the context of cities, using both the central users approach and the geotagged approach was successful as was the application of the LDA, indicating the most proeminent topics, which made it possible to verify the themes most discussed by netizens. The main topics identified through the LDA were related to: Pandemic, with data collected in January 2022, in the first model test, and Urban Mobility, with data collected from September to December 2022, in the second model test. Carrying out the thematic collection on these subjects, the predominant sentiment within the data was negative. It was also found that the supervised strategy was more successful for classifying the sentiment polarity of tweets, especially the SVM and Random Forest methods, especially in the second application test of the model. Having the contribution of the literature and subsequent performance of two tests, the final flowchart as well as a generic model are presented at the end, being, subsequently, raised limitations of the work and opportunities for future studies.

Keywords: Data extraction; Smart Cities; Intermediate Cities; Topic Modeling; Sentiment Analysis; Machine Learning; Twitter

LISTA DE ABREVIATURAS, SIGLAS E ACRÔNIMOS

API	Application Programming Interface
AUC	*Area Under the Curve*
CA	*Classification Accuracy*
CDSTCU	*City Data Scraper Through Central Users*
CGI.BR	Comitê Gestor Da Internet No Brasil
CUP	*Capture, Understand, Present*
ECCO	*Evaluating a Corpus Characterised by Opinion-Bearing Language*
EUA	Estados Unidos da America
IBGE	Instituto Brasileiro de Geografia E Estatística
IDH	Índice de Desenvolvimento Humano
IFDM	Índice Firjan de Desenvolvimento Municipal
IoT	*Internet Of Things*
LDA	*Latent Dirichlet Allocation*
LDAvis	*Interactive Visualization of Topic Models*
MDS	*Multidimensional Scaling*
PB	Portfólio Bibliográfico
PIB	Produto Interno Bruno
PIBIC	Programa Institucional de Bolsas de Iniciação Científica
PPGEPS	Programa de Pós-Graduação Em Engenharia De Produção E Sistemas
PROKNOW-C	*Knowledge Development Process-Constructivist*
RF	*Random Forest*
RSL	Revisão Sistemática De Literatura
RTs	*Retweets*
SVM	*Support Vector Machines*
TIC	Tecnologia Da Informação E Comunicação
UEM	Universidade Estadual De Maringá
UN	Nações Unidas
VADER	*Valence Aware Dictionary and Sentiment Reasoner*

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Este capítulo de Introdução à Dissertação está dividido em subseções, partindo da primeira subseção, da Contextualização, passando pelo Problema de Pesquisa, na sequência, bem como aos Objetivos e Justificativa. Inicia-se com a Contextualização, no subitem abaixo.

1.1 Contextualização

Ao longo das últimas décadas, se observa constante crescimento urbano. Tal observação encontra esteio no relatório de 2018 das Nações Unidas (UN), onde é feito o apontamento do crescimento populacional urbano. Em perspectiva com a década de 50, onde a população urbana correspondia a apenas 30% da população global, sendo que, atualmente, a proporção subiu à casa dos 55%, há a previsão de que, até o ano de 2050, 68% das pessoas estejam residindo nas cidades, globalmente (UN, 2018).

E com o aludido crescimento das cidades, novos desafios surgem em medida proporcional. Uma solução adotada para lidar com os problemas e desafios trazidos pela alta densidade populacional urbana, tem sido a adoção dos conceitos de *smart cities* (MENDONÇA et al., 2016). As *smart cities* – ou cidades inteligentes, em tradução literal – são conceituadas enquanto cidades que possuem ótimo desempenho – e buscam constante aprimoramento – em seis dimensões, levantadas por Giffinger et al., (2007): economia inteligente, pessoas inteligentes, governança inteligente, mobilidade inteligente, meio-ambiente inteligente e vida inteligente.

Avançando no conceito de *smart cities*, outra definição contempla as cidades inteligentes enquanto aquelas em que se verifica a união entre as tradicionais infraestruturas físicas, com as novas tecnologias, de maneira integrada e coordenada (BATTY et al., 2012). Um pressuposto de cidades inteligentes, é, justamente, a promoção de novas formas de engajamento e participação comunitária, por meio das novas tecnologias de informação e comunicação (TIC), de modo a se utilizar destas enquanto novas fontes de dados para planejamento e formulação de políticas (BATTY et al., 2012). Tal pressuposto intersecta uma das seis dimensões de cidades inteligentes propostas por Giffinger et al., (2007), que seria a governança inteligente

(*Smart Governance)*, cuja compreensão abarca, dentre outros aspectos, a disponibilização de serviços aos cidadãos e participação política.

Nesta mesma linha, entende-se que a inteligência das cidades esteja pautada, em grande medida, na integração efetiva entre as “redes de telecomunicações digitais (os nervos), inteligência onipresente (os cérebros), sensores e tags (os órgãos sensoriais) e *software* (o conhecimento e a competência cognitiva)” (CHOURABI et al., 2012, p.2290).

Isto posto, um subgrupo destas ferramentas de TIC é justamente o das Redes Sociais digitais – que neste trabalho serão apenas tratadas como Redes Sociais – as quais, mesmo possuindo diferenças entre si, usualmente, permitem aos usuários “1) construir um perfil público ou semi-público [...], 2) articular uma lista composta de outros usuários com os quais eles compartilham uma conexão e 3) visualizar e percorrer sua lista de conexões e aquelas feitas por outras pessoas, dentro do sistema” (BOYD; ELLISON, 2007, p.211).

Por característica, os dados de Redes Sociais são bastante variados, desde conteúdo textual, audiovisual até metadados, que são produzidos em alta velocidade, ininterruptamente, e em grande volume, o que condiz com os 3 “Vs” característicos da *Big Data* (WANG et al., 2016). A produção de dados nas Redes Sociais conta com ampla difusão e disseminação (LU et al., 2018, KANKANAMGE et al., 2020), sendo ao mesmo tempo um desafio e uma oportunidade, para diversas organizações (tanto públicas quanto privadas) atribuir significado a estes dados, de modo a serem úteis à tomada de decisão (ALSAEDI; BURNAP; RANA, 2017).

Não só sob o ponto de vista prático, como mencionado logo acima, mas o uso de dados de Redes Sociais tem fomentado a criação de novos campos de pesquisa, no qual pesquisadores de variadas áreas tem se debruçado à proposição de novos ângulos e abordagens a problemas como o comportamento urbano, as necessidades da população e suas opiniões, *marketing*, campanhas políticas e outros (SDOUKOPOULOS et al., 2018).

Voltado ao contexto das cidades, a literatura tem abordado a temática das redes sociais de diferentes formas. Há trabalhos com enfoque na proposição de soluções para eventos que ocorrem nas cidades (POORAZIZI; HUNTER; STEINIGER, 2015, ALKHATIB; EL BARACHI; SHAALAN, 2019, ALSAEDI; BURNAP; RANA, 2017, ANDREWS et al., 2016, SÁNCHEZ-ÁVILA et al., 2020, CRAGLIA; OSTERMANN; SPINSANTI, 2012, JAIN; KUMAR, 2018, COSTA et al., 2018, FAN;

JIANG; MOSTAFAVI, 2020, WANG et al., 2020), como desastres, emergências e incidentes de um modo geral, assim como são, também, encontrados trabalhos com foco mais diversificado, propondo novos processos que façam uso de dados de Redes Sociais, para a consecução de objetivos, a exemplo do mapeamento de rotas e comportamentos de grupos usuários (HASNAT; HASAN, 2018, MORA et al., 2018, MUSTO et al., 2015, ABDUL-RAHMAN et al., 2021), tendo ainda aqueles que optam por um enfoque maior à opinião e sentimento públicos (ALKHATIB et al., 2020, MUSTO et al., 2015, ABDUL-RAHMAN et al., 2021, JOSEPH et al., 2017, EL-DIRABY; SHALABY; HOSSEINI, 2019, ALIZADEH; SARLAR; BIRGPUME, 2019, SDOUKOPOULOS et al., 2018, LI et al., 2020, LIU; TENG; GONG, 2021, ADAMU et al., 2021).

Delimita-se, desde já, a opção de que este trabalho irá fazer uso, de maneira mais contundente, deste último grupo de trabalhos. Enquanto *locus* deste trabalho, desde já, define-se a cidade paranaense de Maringá, a qual deteve a posição de número 25 (enquanto cidade mais inteligente do Brasil) no *ranking Connected Smart Cities* – feito pelas empresas *Necta* e *Urban Systems* – sendo, com exceção da capital Curitiba, a cidade melhor colocada, dentre as paranaeneses. Ademais, é a 9ª cidade mais bem colocada do *ranking*, considerando cidades com porte populacional entre 100 mil e 500 mil habitantes[1].

Tendo sido juntado um portfólio de trabalhos que tratam da utilização dos dados de Redes Sociais, passa-se, na sequência, às oportunidades de pesquisa verificadas na literatura, exploradas no subitem da problemática.

1.2 Problema de Pesquisa

Entende-se que a produção de informações úteis à tomada de decisão, tendo por fonte de dados, as Redes Sociais, ainda é questão aberta na literatura, não sendo, esta, uma tarefa trivial, demandando novas soluções (ALKHATIB et al., 2020, ALKHATIB; EL BARACHI; SHAALAN, 2019). Alia-se isso ao fato de que um desafio de inteligência, nas cidades, diz respeito justamente à capacidade de coletar e analisar dados a respeito das pessoas, eventos, estruturas, de maneira tempestiva e

[1] Acesso em 17 de Janeiro de 2022. Disponível em: http://www2.maringa.pr.gov.br/site/noticias/2021/09/02/ranking-aponta-maringa-como-a-cidade-mais-inteligente-do-pr-com-populacao-entre-100-mil-e-500-mil-habitantes/38331.

sustentável (ALKHATIB et al., 2020), de modo a poderem, os gestores, fazer uso do grande volume de dados gerados em Rede Sociais, para melhorar suas decisões e planejamento, e tem-se, assim, uma avenida de pesquisa a ser explorada.

Assim, de maneira mais minuciosa, parte-se, na sequência, para a discriminação de oportunidades para pesquisa verificadas na literatura, a serem consideradas neste trabalho.

1.2.1 Realizar Estudos em Outros Idiomas/Localidades

A partir da Revisão de Literatura realizada, observou-se que a maior parte dos trabalhos que tratavam de analisar dados de Redes Sociais, dentro do amplo escopo comentado, eram predominantemente escritos em língua inglesa, assim como utilizavam-se de fonte de dados textual em inglês. Deste modo, abordagens que utilizem localidades distintas e idiomas distintos, a exemplo do que fizeram Alkhatib, El Barachi e Shaalan (2019) – os quais fizeram uso de *feeds* em árabe, de redes sociais, com foco na identificação de incidentes – têm o condão de enriquecer a literatura e promover a pesquisa sob uma ótica idiomática distinta.

Na mesma linha de Alkhatib, El Barachi e Shaalan (2019), Sánchez-Ávila et al., (2019), em seu trabalho que teve como lócus a Espanha, propuseram um sistema que processava *tweets* em espanhol, acerca de eventos de mobilidade, havendo a recomendação/observação, da aplicação do modelo em outras localidades e idiomas, fazendo adaptações.

Alizadeh, Sarlar e Birgpume (2019), em seu trabalho que tratou do desenvolvimento de duas *dashboards* que capturavam e analisavam dados de várias fontes, inclusive redes sociais, destacaram a participação dos governos locais na consecução do trabalho, sobretudo quando recursos para consecução de engajamento com o cidadão são escassos. Tal pensamento intersecta a ideia de que soluções tecnológicas que façam uso de dados de redes sociais, cujo custo é menos pesado ao erário público do que sensores físicos, por exemplo, são sobretudo importantes a localidades menos desenvolvidas ou em desenvolvimento (LU et al.,

2018), como é o caso do Brasil – no geral – país em desenvolvimento[2], lócus amplo deste estudo.

Deste modo, enquanto lócus, delimita-se, incialmente, o Estado do Paraná, mais especificamente, as *Smart Cities* paranaenses. O Paraná é o segundo estado brasileiro com o maior número de cidades inteligentes, com nove (atrás apenas de Minas Gerais), de acordo com ranking *Connected Smart Cities* 2021, realizado pela empresa *Urban Systems*[3].

As nove cidades paranaenses, classificadas como inteligentes, de acordo com o *ranking* mencionado, são: Curitiba, Maringá, Londrina, Apucarana, Foz do Iguaçu, Cascavel, Pato Branco, Pinhais e Toledo. Destas, somente Curitiba é que não pode, sob nenhum critério, ser classificada enquanto cidade média ou intermediária. As demais, todas, são lócus potenciais deste estudo, bem como outras cidades brasileiras, inteligentes, intermediárias.

Optou-se pela escolha da cidade de Maringá, enquanto lócus único deste trabalho. Das razões que conduziram à sua escolha, duas se destacam. Maringá, dentre todas as cidades paranaenses de porte intermediário, sob a ótica do tamanho da população (entre 100 e 500 mil habitantes), foi apontada como a mais inteligente, pelo *Connected Smart Cities*. Além disso, foi aquela em que os dados foram obtidos em maior volume (por razões que serão mais bem fundamentadas na Metodologia), considerando o curto período havido para coleta, fato que consolidou sua escolha.

1.2.1.1 Cidades Intermediárias: definição e trabalhos

Conceitualmente, o entendimento acerca do que significa rotular uma cidade enquanto média ou intermediária não pode ser pautado em uma visão que considere uma variável, apenas, a exemplo do tamanho demográfico. Entende-se que tal análise deve se pautar em uma miríade de critérios que englobam, além do tamanho demográfico, fatores como infraestrutura da cidade, distância de áreas metropolitanas, as funções que a cidade desempenha e oferta de bens e serviços, capacidade de retenção da população, relações interurbanas e com o campo e indicadores de

[2] Acesso em 27 de Janeiro de 2022. Disponível em: https://www.un.org/development/desa/dpad/publication/world-economic-situation-and-prospects-february-2020-briefing-no-134/.
[3] Acesso em 1 de Março de 2022. Disponível em: https://ranking.connectedsmartcities.com.br/.

qualidade de vida, por exemplo (SILVA, 2013). Silva (2013) aponta, ainda, que uma cidade média realiza intermediação entre os espaços locais e os regionais.

Maringá, cidade planejada, de porte médio (RODRIGUES, 2004, SAVI; CORDOVIL, 2015), localizada no norte do estado do Paraná, tem população estimada em 436.472 pessoas[4]. Possui, segundo IBGE, classificação de Capital Regional B, na hierarquia urbana, dada a cidades que são “centros de referência no interior dos estados”. As capitais regionais, no geral, são “centros urbanos com alta concentração de atividades de gestão, mas com alcance menor em termos de região de influência”[5]. Portanto, Maringá não se confunde com uma metrópole, mas, pelo contrário, exerce influência central na região onde está localizada, no interior do estado do Paraná, seja por seu volume demográfico, oferta de estruturas de bens e serviços, ou outros critérios trazidos.

Rumando à produção acadêmica, sobretudo àquela com propósitos parecidos ao que aqui se intenta, reforça-se que não foi encontrado, na literatura revisada por pares, trabalhos em português, com objetivos similares ao que aqui se propõe, independente do porte da cidade lócus.

Verifica-se, na literatura, que trabalhos que tenham como lócus grandes centros urbanos, sobretudo tendo em vista um aspecto mais voltado ao volume demográfico, observam uma facilidade maior na extração de dados em grandes volumes, tendo em vista a maior quantidade de *netizens* (cidadãos internautas) destas localidades. Trabalhos como os de Andrews et al., (2016) e Wang et al., (2020) que tiveram como lócus localidades como *Nepal* e *Houston* tiveram *dataset*s (iniciais) de tamanho que variou entre 984.643 e 7.041.794 *tweets*, coletados num período de 3 e 34 dias, respectivamente.

Ainda, trabalhos com pesquisa meramente temática, como fizeram Gasco et al., (2019) em seu trabalho sobre monitoramento de barulho, conseguem obter, mais facilmente, *datasets* volumosos. No caso de Gasco et al., (2019) foi obtido um *dataset* inicial de mais de 5,6 milhões de *tweets,* num período de coleta que compreendeu 3 meses, não tendo havido qualquer restrição de geolocalização. Quanto a este último

[4] Acesso em 27 de Janeiro de 2022. Disponível em: https://cidades.ibge.gov.br/brasil/pr/maringa/panorama.

[5] Acesso em 27 de Janeiro de 2022. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101728_folder.pdf.

ponto, Gasco et al., (2019) reconhecem vantagens de “setar” parâmetros de localização para coleta de *tweets,* e recomendam tal escolha para trabalhos futuros.

Ademais, necessário levar em conta a utilização do *Twitter* em diferentes países. Verifica-se, em dados de outubro de 2021[6], que os Estados Unidos (EUA) é o país com maior número de usuários, com 77,75 milhões, sendo que o Brasil aparece na 4ª posição, com um número de usuários de 19,05 milhões. Proporcionalmente, a população norte-americana é 55% maior do que a brasileira, (329,5 e 212,6 milhões, respectivamente), contudo, a proporção de usuários do *Twitter* entre estes dois países não segue a mesma linearidade, tendo os EUA uma população de *netizens,* na citada rede, superior em mais de 300% à brasileira. Tal dado sustenta o fato de que, mesmo trabalhos com lócus situados em cidades menores, americanas, percebem maior facilidade na obtenção de um conjunto de dados maior. Exemplo disso é o trabalho de Yuan et al., (2021), com lócus na cidade americana de Wilmington, localizada no estado de North Carolina, com população de 120.194 habilitantes[7], ter obtido uma *dataset* de mais de 2,6 milhões de *tweets* geolocalizados, num período de 11 dias.

Deste modo, entende-se haver um complicador inerente à localização de trabalhos, no contexto específico de uma ou mais cidades, sobretudo quando se está lidando com cidades brasileiras, que não são grandes metrópoles, com um volume menor de *netizens* no *Twitter,* por consequência. Assim, acessar os principais tópicos – relacionados ao contexto urbano – sejam problemas, reclamações, situações, angústias ou até mesmo elogios, para posterior análise de sentimento focada, de maneira geograficamente localizada, é tarefa que, diante do contexto que aqui se propõe, requer novas soluções.

1.2.2 Usar de Ferramentas livres num contexto de Cidades Inteligentes

Ainda, para consecução dos objetivos propostos aqui, serão utilizadas ferramentas tais como *Python 3.0* e *Orange Data Mining* – para acesso e análise dos dados da Rede Social *Twitter* – ambas ferramentas livres, do ponto de vista da distribuição e acesso. A utilização de ferramentas livres é uma das proposições da visão de *smart city*, que vai ao encontro da adesão ao movimento de utilização de

[6] Acesso em 1 de Março de 2022. Disponível em: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/.
[7] Acesso em 1 de Março de 2022. Disponível em: https://www.census.gov/.

softwares livre e *open access* (de acesso aberto) (ALIZADEH; SARLAR; BIRGPUME, 2019).

Sdoukopoulos et al., (2018) fizeram uso de outra solução do tipo livre (*Microsoft Public License*), para acessar *tweets* sobre o transporte público de Londres e realizar análise de sentimentos, dentre outras, chamada *NodeXL*, o que consona com a ideia da utilização de *softwares* livres, tendo ainda a vantagem de que, assim como o *Orange Data Mining*, também não requer habilidades de programação em código (SDOUKOPOULOS et al., 2018).

Retornando à recomendação da utilização de algoritmos baseados em *Python*, observa-se sua adesão na literatura acadêmica (ALKHATIB; EL BARACHI; SHAALAN, 2019, YUAN et al., 2021, ABDUL-RAHMAN et al., 2021, LI et al., 2020). Acredita-se que as razões para a utilização de *Python* se devam ao fato de ser *open source*, uma linguagem simples de aprender, e barata (ABDUL-RAHMAN et al., 2021).

1.2.3 Agregar outras análises

Outra recomendação proeminente que se observou nos trabalhos levantados, é a de se agregar diferentes tipos de análise num mesmo trabalho, a exemplo de que se intenta fazer análise de sentimentos, fazer a consideração, também, do aspecto temporal presente no *dataset*, ou seja, da evolução do sentimento ou opinião pública ao longo de um período (ANDREWS et al., 2016, GASCO et al., 2019, ALKHATIB et al., 2020).

Ainda, alguns dos trabalhos, conforme já comentado, tem uma abordagem focada em detecção/identificação de eventos, na medida em que, neste caso, uma análise recomendada a se fazer, para além da mera identificação do incidente, é a análise do sentimento das pessoas, no momento em que o evento ocorre (ALKHATIB; EL BARACHI; SHAALAN, 2019).

Evoluindo na consideração do aspecto temporal, nas análises, outra recomendação observada, é a da análise de tendência, a exemplo do que pode ser feito tendo em mãos a data de criação de *tweets*, por exemplo, de modo não só a verificar a evolução de um tema ao longo do tempo, mas identificar tendências (MARZOUKI et al., 2021).

Para além de análise de tendências, também se aponta para pesquisas com análises mais longitudinais, utilizando períodos maiores, de modo a observar

mudanças/variações nos sentimentos, fazendo isso, inclusive, de maneira *multilocus*, de modo a permitir a comparabilidade de diferentes localidades, além de proporcionar maior riqueza analítica (ALIZADEH; SARLAR; BIRGPUME, 2019).

Outra recomendação é a de tentar aprofundar as análises feitas, considerando diferentes *features* do conteúdo extraído – diferentes atributos de um *tweet* – como também o escopo geográfico da postagem, *hyperlinks*, dentre outros (EL-DIRABY; SHALABY; HOSSEINI, 2019, MARZOUKI et al., 2021), para que a análise tenha maior peso e significado.

Referente à inclusão da dimensão da localidade, deve ser dito que, existem limitações no volume de dados possíveis de se obter tende em vista a maior parte dos *tweets* serem postados sem que o usuário compartilhe sua localização exata (haja vista esta ser uma opção de fazê-lo ou não) (SDOUKOPOULOS et al., 2018, SÁNCHEZ-ÁVILA et al., 2020), a exemplo de trabalhos como o de Yuan et al., (2021), onde menos de 3 (três) por cento do total de tweets do *dataset* era geolocalizado.

Assim, de modo a tentar contornar tal limitação, buscar-se-á a extração de dados levando em conta Usuários Centrais, conforme comentado na sequência.

1.2.4 Identificar e Utilizar dados oriundos de Usuários Centrais

Outra recomendação que vai ao rumo de análises mais sustentáveis, com *datasets,* por vezes, menores, mas mais representativos da população, é a de utilizar conteúdo “amostral” significativo para o que se pretende analisar. Tal conteúdo é chamado de maneira distinta a depender do trabalho que se está fazendo, podendo ser Usuários Centrais, Líderes de Opinião ou Formadores de Opinião, traduções aproximadas para *Central Users, Opinion Leaders* e *Influential Spreaders*, no inglês (SDOUKOPOULOS et al., 2018, ALKHATIB et al., 2020, ABDUL-RAHMAN et al., 2021, YUAN et al.,2020, LI et al., 2020, ADAMU et al., 2021).

Apesar de se ter claro que uma amostra maior – como foi a recomendação ao fim dos trabalhos de Hasnat e Hasan (2018) e Adamu et al., (2021) – em teoria, traduz análises que sejam mais representativas de determinados fenômenos, nem sempre tal escolha é assertiva, na medida em que se deve levar em conta o tempo de processamento de uma quantidade maior de dados, que vai de encontro à necessária sustentabilidade do método, além de que tal lógica esbarra nas próprias

restrições/limitações do volume alcançável, utilizando APIs gratuitas para acessar dados do *backend* do *Twitter*.

Neste sentido, argumenta-se, ainda, a respeito da falta de escalabilidade de métodos que sejam dependentes de um *dataset* muito volumoso, o que, por sua vez, implica em ineficiência e demanda um tempo para processamento e análise maiores (ALKHATIB et al., 2020).

Ainda, Alkhatib et al., (2020) argumentam que o uso de conteúdo de alto impacto, como publicações de líderes de opinião, é bastante útil e uma maneira sustentável para análise do sentimento e opinião públicos. Em seu trabalho, Alkhatib et al., (2020) propuseram uma *framework* para monitoramento da opinião pública, tendo por novidade, o foco está nos *posts* de alto impacto de líderes públicos de opinião e seus seguidores, de modo a fornecer entendimento em profundidade e gerar relatórios do sentimento público.

Na mesma linha, Abdul-Rahman et al., (2021) destacaram como recomendações para trabalhos futuros, a identificação de *influential spreaders* dentro do *dataset*, e fazer a focalização nestes de modo a obtenção de informações.

Recomendação no mesmo tom é dada por Yuan et al., (2020), onde os autores ratificam a importância na identificação de grupos de usuários alvo, considerando dados de redes sociais de modo a ter um direcionamento maior na pesquisa, e um delineamento mais bem definido.

Ainda, de um modo geral, acredita-se que os dados de redes sociais (como um todo) sejam cada vez mais representativos da realidade a qual se deseja aceder no estudo, com o passar do tempo (SDOUKOPOULOS et al., 2018). Isso dá margem para que a utilização de Usuários Centrais como fonte de dados, cresça, também em representatividade da População.

A ideia da utilização de Usuários Centrais é aqui defendida, na medida em que utilizar-se-á de duas redes oficiais municipais no *Twitter* para a coleta de dados, sob as quais serão realizadas a extração e análise dos dados. Trata-se das contas oficiais da Prefeitura da Cidade de Maringá, bem como a do Prefeito de Maringá. Maiores informações e fundamentações serão trazidas na seção da metodologia, mais adiante.

1.2.5 Questão de Pesquisa

Diante das razões fundamentadoras trazidas com a revisão da literatura, nos tópicos levantados nesta seção, definem-se, enquanto Tema, Delimitação e Questão de Pesquisa, os itens abaixo:

- TEMA: Extração e análise de dados de Redes Sociais
- DELIMITAÇÃO: Extração e análise de dados de Redes Sociais em cidades intermediárias, brasileiras.
- QUESTÃO DE PESQUISA: Como extrair dados da Rede Social *Twitter*, geograficamente situados em cidades médias brasileiras, para análise de sentimentos?

1.3 Objetivos

Assim, o que se propõe aqui é, por meio da aplicação da Alocação Latente de Dirchlet (LDA), identificar o(s) tópico(s) proeminente(s) discutidos por *netizens* de uma cidade inteligente intermediária paranaense, e realizar análise de sentimentos deste, tendo como fonte de dados a rede social *Twitter*.

Intenta-se, também, a obtenção dos dados de conteúdo relevante, advindo de Usuários Centrais, utilizando-se, neste caso de duas contas oficiais municipais, referentes à cidade lócus, no *Twitter,* para a coleta dos dados, sob os quais serão aplicadas as análises propostas. Trata-se das contas oficiais da Prefeitura da Cidade de Maringá, bem como a do Prefeito de Maringá. Maiores informações e fundamentações serão trazidas na seção da metodologia, mais adiante. De uma forma mais canônica, são apresentados os Objetivos, nas subseções a seguir.

1.3.1 Objetivo Geral

propor e testar um modelo para extração e análise de dados/opinião pública, situado no contexto das cidades intermediárias, tendo por fonte a rede social *Twitter*.

1.3.2 Objetivos Específicos:

a) Elaborar código para extração de dados utilizando a abordagem de usuários centrais, por meio da linguagem aberta de programação *Python*;
b) Identificar o(s) tópico(s) proeminente(s), emergido(s) no conjunto de dados extraídos;

c) Realizar a extração de dados temática, na sequência, com base no(s) principal(is) tópico(s) identificado(s), utilizando a abordagem de usuários centrais, e a abordagem geolocalizada;
d) Aplicar modelos de Análise de Sentimento, no conjunto de dados temáticos extraídos, referente ao(s) principal(is) tópico(s);

1.4 Justificativa

A produção de informações úteis à tomada de decisão, tendo por fonte de dados, as Redes Sociais, a exemplo do *Twitter*, ainda é questão aberta na literatura, não sendo, esta, uma tarefa trivial, demandando novas soluções (MUSTO et al., 2015, ANDREWS et al., 2016, ALSAEDI; BURNAP; RANA, 2017, ALKHATIB et al., 2020).

São trazidas, nesta subseção da Justificativa, as razões, sobretudo teóricas e práticas, que levam ao entendimento da relevância inequívoca da presente dissertação.

1.4.1 Justificativa Teórica

Inicia-se a justificativa para realização deste trabalho, do ponto de vista teórico, trazendo intersecções, justamente, com a razão "geográfica" para realização da dissertação, qual seja, ter sido observada na literatura a predominância de estudos em localidades que tem como lócus a língua inglesa. Estudos tais como o de Alkhatib, El Barachi e Shaalan (2019) – que fizeram a utilização de *feeds* de redes sociais em linguagem árabe – são exceção, e o fato de modelos de *unsupervised machine learning* (Aprendizado de Máquina Não Supervisionado), tal qual o VADER[8], para Análise de Sentimentos, ser focado em conteúdo em inglês, já traz um nível a mais de complexidade analítica.

Abrindo parêntesis, Análise de Sentimento se define como a categorização das opiniões de um grupo diante de um dado assunto, sendo que sua análise é feita com base em conteúdos textuais, para os quais as redes sociais são sólidas fontes de dados (USHARANI, 2018), podendo ser processada, pelos métodos de aprendizado de máquina de maneira não supervisionada, como o método VADER

[8] Acesso em 10 de Agosto de 2022. Disponível em: https://github.com/cjhutto/vaderSentiment

citado acima, ou supervisionada. Reforça-se que o fato de o VADER ser focado em conteúdo na língua inglesa acrescenta um complicador a mais, fazendo com que a utilização da estratégia Supervisionada seja necessária, neste caso, a qual será tratada mais adiante, na seção da metodologia.

Ainda, também do ponto de vista teórico, intersectando o *locus* de pesquisa, entende-se haver um complicador inerente à utilização de uma cidade brasileira como estudo de caso, não só em função da linguagem dos *tweets,* os quais, aqui, serão, logicamente, em português, mas também por se tratar de cidade de porte intermediário, conforme já comentado e fundamentado na seção da problemática.

Contudo, a consecução dos objetivos aqui propostos pode ser valorosa a cidades com porte e estruturas semelhantes à Maringá, *locus* deste trabalho, isto porque, a abordagem aqui proposta pode ser replicada em outros casos, com as devidas pequenas adaptações, quanto à seleção das contas dos usuários centrais, considerando sua disponibilidade na rede social *Twitter*.

Maringá, tal qual já comentado, no *ranking Connected Smart Cities* – feito pelas empresas *Necta* e *Urban Systems* – deteve a posição de número 20 enquanto cidade mais inteligente do Brasil, sendo a 7ª cidade mais bem colocada do *ranking*, considerando cidades com porte populacional entre 100 mil e 500 mil habitantes[9] (era a cidade de número 25 e de número 9, nos critérios elencados, até então, segundo o *ranking* anterior[10]). Levando em conta as 20 primeiras cidades classificadas no *ranking,* dentro do estrato que compreende cidades com populações entre 100 e 500 mil habitantes, verificou-se que 13 destas possuem, assim como Maringá, a classificação, pelo IBGE, enquanto Capital Regional, representando, assim, potenciais *locus* de estudo.

Justifica-se, ainda, em razão da sustentabilidade da abordagem aqui proposta, a qual utiliza-se de usuários centrais para coleta de dados, segundo a qual, o *dataset* conseguido é menor, do que caso fosse optada pela coleta ampla, utilizando apenas *Keywords*, mas é, ao mesmo tempo, mais representativo dos anseios da população *locus* do estudo, usuária da rede social *Twitter*, que se expressa, neste

---

[9] Acesso em 28 de Fevereiro de 2022. Disponível em: http://www.maringa.pr.gov.br/site/noticias/2022/11/01/maringa-e-a-7-cidade-mais-inteligente-do-brasil-entre-os-municipios-de-100-mil-a-500-mil-habitantes-aponta-ranking/40605

[10] Acesso em 28 de Fevereiro de 2022. Disponível em: http://www2.maringa.pr.gov.br/site/noticias/2021/09/02/ranking-aponta-maringa-como-a-cidade-mais-inteligente-do-pr-com-populacao-entre-100-mil-e-500-mil-habitantes/38331.

caso, diretamente aos tomadores de decisão. Tal abordagem é reputada como mais sustentável, na medida em que, tendo um conjunto de dados menor, mas mais representativo, o ruído é menor, com menos dados “inúteis” à análise que aqui se propõe, o que, por consequência, permite a escalabilidade desta abordagem, tendo em vista um tempo de processamento menor (SDOUKOPOULOS et al., 2018, ALKHATIB et al., 2020, ABDUL-RAHMAN et al., 2021, YUAN et al.,2020, LI et al., 2020, ADAMU et al., 2021).

Para tanto, seguir-se-á também o que preconiza a literatura, no que diz respeito à utilização de *softwares* livres, num contexto de cidades inteligentes, onde, aqui, implementar-se-á as estratégias propostas por meio de programação em *Python* e pelo *software Orange Data Mining*, o qual, por sua vez, também é baseado em *Python*, linguagem que, no geral, encontra esteio no portfólio de artigos revisados (ALKHATIB; EL BARACHI; SHAALAN, 2019, YUAN et al., 2021, ABDUL-RAHMAN et al., 2021, LI et al., 2020).

De maneira sintetizada, considerando o objetivo de proposição de um modelo para extração de dados para análise de sentimentos, tendo por fonte a rede social *Twitter*, num contexto de cidade, são trazidos, abaixo, alguns dos principais modelos observados na literatura revisada, e limitações observadas, frente ao modelo que, adiante, será apresentado:

Quadro 1 - Modelos da Literatura

Trabalho	Características	Limitações
ABDUL-RAHMAN et al., 2021	Propõe framework para mineração e análise de dados, geolocalizados, do Twitter. Utilização de técnicas de Processamento de Linguagem Natural e Modelagem de Tópicos (LDA), no teste do modelo, em um estudo de caso. Fonte de dados: *Twitter*	necessidade de haver *tweets* geolocalizados; *dataset* compreendendo um período muito longo; falta de escalabilidade
ADAMU et al., 2021	Parte de uma plataforma já conhecida (CRISP-DM) para propor uma análise de emoções, tendo por base conteúdo textual em uma variação informal do inglês (Pidgin), construindo, para tanto, um *emotion dataset* nesta língua. O tema foi pré-definido e concerneu à percepção do sentimento público frente à atuação do Governo Nigeriano na	muito abrangente (nacional); não contempla nível de cidade; temática pré-estabelecida

	distribuição de paliativos contra a COVID-19. Fonte de dados: *Twitter*	
ALKHATIB et al., 2020	Propõe framework para monitoramento da opinião pública em tempo real com base nos posts de alto impacto de líderes públicos de opinião e seus seguidores, de modo a fornecer entendimento em profundidade e gerar relatórios do sentimento público. Fonte de dados: *Twitter*	muito abrangente (nacional); não contempla nível de cidade; temática pré-estabelecida; temática pré-estabelecida
ALKHATIB; EL BARACHI; SHAALAN, 2019	Propõe *framework* para gestão de eventos e incidentes em cidades inteligentes, a qual tem o objetivo de fornecer relatórios em tempo real sobre eventos e incidentes, de modo a prover os órgãos de resposta a emergência com informações úteis para planejamento e tomada de decisão. Fonte de dados: *Twitter*	*dataset* compreendendo um período muito longo; não tem foco em análise de sentimentos
EL-DIRABY; SHALABY; HOSSEINI, 2019	Propõe metodologia para avaliar a satisfação dos usuários de transporte público, por meio na análise de redes sociais (*social network analysis*), detecção de tópicos nos dados e análise de sentimento nos tópicos. Fonte de dados:	não contempla nível de cidade; temática pré-estabelecida
FAN; JIANG; MOSTAFAVI, 2020	Propõe *framework* para detecção de disrupções/incidentes/emergências nas cidades, com base na frequência de tweets sobre dado assunto, combinando, junto da análise textual, a análise das descrições de imagens tuitadas e analise geográfica.	necessidade de haver *tweets* geolocalizados para análise geográfica; abordagem aplicável a grandes cidades – grandes volumes de dados.
LI et al., 2020	Constrói um programa baseado em *Python* para extrair opiniões de Rede Social sobre modificação genética em alimentos, identificando tópicos e aplicando análise de sentimentos por meio da regressão logística multivariada. Fonte de dados: Sina Weibo (*Twitter chinês*)	*dataset* compreendendo um período muito longo; não contempla nível de cidade; temática pré-estabelecida
MUSTO et al., 2015	Propõe *framework* para análise textual, em tempo real, do conteúdo de redes sociais, aplicando-se, também, análise de sentimentos e plotagem em mapa,	abordagem melhor aplicável a grandes regiões – grandes volumes de dados (a exemplo do teste de

		aplicação do modelo *italian heat map,* feito a nível de país),

Fonte: dados de pesquisa (2022)

Desde já se argumenta que o modelo que será apresentado neste trabalho contempla uma abordagem para coleta de dados baseada em Usuários Centrais, por meio da qual é possível situar a coleta no contexto da cidade, de uma maneira sustentável, escalável, para posterior análise de dados.

Tendo sido cumprida a subseção de Justificativa Teórica, passa-se, na sequência, à apresentação da Justificativa Prática.

1.4.2 Justificativa Prática

Justifica-se, na prática, a realização desse trabalho, enquanto um novo meio a ser utilizado pelos gestores públicos para que ouçam os anseios da população das cidades das quais são responsáveis. Por meio do modelo que aqui se propõe, os gestores poderão ganhar *insights* tanto sobre os principais tópicos que estão sendo discutidos no momento, pelos *netizens*, assim como o sentimento das pessoas sobre estes tópicos. Na literatura, se entende a opinião pública enquanto a intersecção entre um tópico de interesse em um conjunto de dados, e o sentimento da fonte produtora do dado, sobre aquele tópico (EL-DIRABY; SHALABY; HOSSEINI, 2019). É esta, também, a definição de opinião pública adotada na presente pesquisa.

Ainda, deve-se ter em mente que abordagens que façam uso de dados de Redes Sociais – como é o caso deste estudo, que faz uso de dados produzidos pelos usuários do *Twitter* – não encontram as limitações temporais e de custo que *surveys* tradicionais percebem (LI et al., 2020), o que as tornam fontes mais céleres e eficientes, podendo, em última instância, servirem de complemento à outras fontes de dados (EL-DIRABY; SHALABY; HOSSEINI, 2019).

Ademais, por se tratar de fonte de dados cuja produção (*tweets*) dos dados pelos *netizens,* usuários da rede social, não é induzida, tem-se o entendimento de que, há o potencial de representarem mais fidedignamente as opiniões e sentimentos públicos de uma dada população (ALIZADEH, SARLAR, BIRGPUME, 2019). Ressalta-se, no entanto, que a amostra pode ser considerada induzida sob a

perspectiva da escolha dos usuários centrais, por meio dos quais se operacionalizará a coleta de dados. Mais sobre a escolha dos Usuários Centrais, no decorrer do trabalho.

O perfil[11] do usuário do *Twitter* é predominantemente masculino, compreendido numa faixa etária de 18 até 49 anos. De todo modo, arrazoa-se, também, no sentido de que cada vez mais seja utilizado o *Twitter* – bem como Redes Sociais, no geral – pelas autoridades públicas, para que seja dada maior transparência nas ações governamentais, mas também para que haja esse engajamento com a comunidade de usuários, tendo em vista, sobretudo, o fato de que a parcela de pessoas presente em redes sociais só tem a crescer com o passar do tempo, tornando cada vez mais representativa, da população, as opiniões ali exaradas (SDOUKOPOULOS et al., 2018). Não só do ponto de vista das pessoas, cada vez mais utilizando redes sociais, mas também organizações públicas cada vez mais presentes, visto que tal fato é corroborado por relatório recente do Comitê Gestor da Internet no Brasil (CGI.br)[12], onde é apontado crescimento na presença de organizações públicas nas redes sociais, de de 82% em 2019 para 94% em 2021.

Ressalta-se, de igual modo, desde logo, que é, justamente, a presença dos gestores e organizações públicas, nas redes sociais – mais especificamente no *Twitter,* neste caso – que garante a potencialidade da abordagem aqui proposta, para extração e análise dos dados, contextualmente situados nas cidades.

Reforça-se, do ponto de vista prático, que as informações obtidas da análise do modelo proposto podem ser úteis à gestão das cidades, pois, num primeiro momento, indica o(s) assunto(s) mais falado(s) pelos cidadãos, de maneira tempestiva (trânsito, investimentos com o orçamento público, por exemplo), podendo revelar descontentamento público com as decisões tomadas, além de ser possível, também a obtenção de *insights*, ou sugestões diretamente das pessoas afetadas pela ação dos tomadores de decisão.

As informações obtidas pela análise, com a aplicação do modelo, podem também servir como indicadores de efetividade da execução de políticas públicas,

---

[11] Acesso em 1 de março de 2023. Disponível em: https://www.agenciamestre.com/redes-sociais/estatisticas-twitter-brasil/

[12] Acesso em 6 de Agosto de 2022. Disponível em: https://www.cgi.br/media/docs/publicacoes/2/20220725170710/tic_governo_eletronico_2021_livro_eletronico.pdf.

revelando, inclusive, a depender do caso, a satisfação das pessoas com ações da Gestão, as quais podem ser reforçadas.

Ainda, a tempestividade da abordagem, muito devida tendo em vista a representatividade e objetividade dos dados tomados para análise, torna possível, não só a correta identificação do(s) principal(is) tópico(s) discutidos pelos *netizens*, como faculta também agir de maneira rápida, evitando o recrudescimento de cenários causadores de sentimento, por exemplo.

Com tudo o que fora dito, só se reforça o alinhamento e atendimento à dimensão de cidade inteligente que trata da governança (GIFFINGER et al., 2007) sobremaneira quanto a levar em conta à opinião pública.

Vencida a subseção de Justificativa Prática, passa-se, na sequência, à apresentação da Justificativa Pessoal.

1.4.3 Justificativa Pessoal

Pessoalmente, a temática das Redes Sociais é cara, ao autor deste trabalho, enquanto assunto de pesquisa, desde o ano de 2017, em que, bolsista do Programa Institucional de Bolsas de Iniciação Cientifica (PIBIC), desenvolveu um trabalho de cunho qualitativo que tratava de analisar as publicações de um órgão de segurança pública, em sua conta no *Facebook*, tendo por bojo, os conceitos de Cidades Inteligentes.

Assim, enquanto aluno da Pós-Graduação em Engenharia de Produção e Sistemas (PPGEPS), a possibilidade de continuar trabalhando nesta linha, mas agora com métodos mais automatizados e escaláveis – em detrimento ao trabalho prévio, em que a análise feita foi manual – é algo apreciável, e que tem suscitado a aquisição de novos conhecimentos, tanto teóricos quanto práticos. Ainda, enquanto servidor público, a proposição de melhoramentos à gestão, de um modo ou de outro, tendo em vista a melhoria do atendimento às necessidades das pessoas, e à administração da *res publica,* também é algo que o autor deste tem estima.

1.5 Estrutura da Dissertação

Esta dissertação, para além deste primeiro capítulo, de Introdução, apresentará, na sequência, um capítulo de Revisão de Literatura, seguido por outro

de Metodologia, cada qual, contendo subdivisões internas. Posteriormente será trazido um capítulo de resultados, de duas execuções de teste de aplicação do modelo, que foram feitas, com o capítulo de conclusões fechando o trabalho.

2 REVISÃO DE LITERATURA

O presente capítulo tem por objetivo trazer à baila os trabalhos da literatura relacionada, que ajudaram a fundamentar as escolhas feitas, pela pesquisa que aqui se propõe.

Primeiramente, serão discutidos os procedimentos e metodologias de revisão sistemática seguidos para obtenção do Portfólio Bibliográfico (PB) de artigos base para o presente trabalho, e, então, na sequência, serão trazidas as colocações conceituais sobre os principais temas que intersectam os objetivos do presente trabalho, nas subseções seguintes. Passa-se, na sequência, à descrição da Revisão Sistemática de Literatura.

2.1 Descrição da Revisão Sistemática de Literatura (RSL)

Para a obtenção do conjunto de artigos que fundamentam, teoricamente, o que aqui se pretende realizar, foi conduzida a revisão sistemática de literatura, através de duas metodologias de revisão distintas, quais sejam: *Proknow*-C (ENSSLIN; ENSSLIN; PINTO, 2013) e Modelagem de Tópicos para Revisão de Literatura – adaptado de Asmussen e Möller (2019) – nesta ordem.

Justifica-se ter sido realizada a revisão em duas etapas, em função de algumas razões. A primeira delas, é a de que, com duas metodologias, consegue-se a “confirmação” da relevância de artigos que porventura surjam em ambas, como, inclusive, é o caso aqui. Trata-se de redundância que subsidia a construção de um portfólio robusto. A segunda razão é que, entre uma e outra aplicação das metodologias de revisão houve um intervalo temporal de mais de seis meses, assim, até mesmo buscando a atualidade do portfólio, é que reputou-se adequada a aplicação da segunda metodologia, por meio da modelagem de tópicos, mesmo já tendo sido realizado do *Proknow*-C. Finalmente, a operacionalização da segunda metodologia de revisão, por meio da Modelagem de Tópicos teve o intento de automatizar e celerizar as primeiras etapas de revisão, retirando o “olhar humano”, falível e que pode, pelo cansaço destas etapas iniciais mais “braçais” acabar tomando decisões equivocadas quanto a manutenção ou não de dado artigo, para as etapas seguintes, o que, uma vez mais, vai ao encontro de se buscar quase que uma ratificação da robustez do portfólio final.

Serão comentadas, abaixo, separadamente, cada uma destas etapas/metodologias de revisão, desde eixos e palavras-chave utilizados para busca, bases de dados utilizadas e o PB final de artigos, ao término de cada metodologia.

2.1.1 Proknow-C

A primeira estratégia para revisão de literatura utilizada foi feita por meio da utilização da técnica de revisão sistemática de literatura denominada *Knowledge Development Process-Constructivist* (*Proknow-C*). O *Proknow-C* compreende um processo estruturado de seleção da literatura científica, com o fulcro de prover ao pesquisador o conhecimento acadêmico produzido sobre determinado tema (ENSSLIN; ENSSLIN; PINTO, 2013).

O processo é composto, a grosso modo, por quatro macro etapas, quais sejam "(a) seleção de um portfólio de artigos sobre o tema da pesquisa; (b) análise bibliométrica do portfólio; (c) análise sistêmica; e, (d) definição da pergunta de pesquisa e objetivo de pesquisa" (ENSSLIN; ENSSLIN; PINTO, 2013, p.333).

No presente trabalho, contudo, explicita-se a operacionalização proposta somente da etapa "(a)", de seleção de um portfólio de artigos sobre o tema da pesquisa, tendo em vista o objetivo primário, de com isso, apresentar o PB de artigos levantado, através dos passos da metodologia *Proknow-C*.

Utilizando-se da metodologia *Proknow*-C, para RSL A pesquisa dos artigos para a seleção do portfólio é feita por meio da definição de eixos temáticos (conforme os assuntos de interesse) e palavras-chaves escolhidas para cada um, contudo, antes disso, deve-se escolher as bases de dados nas quais serão buscados os artigos.

As bases de dados escolhidas foram duas: *Scopus* e *Web of Science*. As razões que fundamentaram a escolha devem-se sobretudo ao fato de serem bases multidisciplinares, que congregam uma variedade grande de periódicos – revisados por pares – e também o fato de ambas admitirem a busca utilizando-se de operadores *booleanos*, o que facilita a operacionalização.

Tendo sido escolhidas as bases de dados, passou-se à definição dos Eixos e Palavras-Chave, que foram, em primeiro momento, os constantes no Quadro 1:

Quadro 2 - Eixos e Palavras-chave *Proknow-C*

EIXOS			
1	**2**	**3**	**4**

Dados de Redes Sociais	Abrangência Cidades	Eventos/ Problemas; Processos/ Operações	Inteligência
"social media" OR "social network" OR "sns" OR "digital media" OR "human sens*" OR "social sens*"*	*"cit*" OR "region*" OR "municip*" OR "govern*" OR "local*" OR "urban*" OR "communit*"*	*"disaster" OR "emergenc*" OR "crisis" OR "problem*"OR "incident*" OR operation* OR process* OR management OR service**	*"smart*" OR "digital" OR "intelligen*" OR "sustainab*" OR "resilien*"*

Fonte: dados de pesquisa (2021)

O Eixo 1 consistiu em palavras-chave que remetessem ao conceito de Dados de Redes Sociais.

O Eixo 2 compreendeu palavras-chave onde constasse delimitado o contexto de trabalho que se está buscando, qual seja: abrangência/contexto de cidades.

O Eixo 3 tratou de abarcar trabalhos que tratassem de problemas, eventos ocorridos nas cidades, e também processos e operações nesse mesmo contexto, ambos, relacionado ao uso de dados de redes sociais.

Por fim, o Eixo 4 tratou de intersectar à inteligência, em função de haver, emoldurando o trabalho, o bojo téorico das *smart cities*.

A definição de palavras-chave, para cada eixo, levou em conta artigos alinhados, previamente fichados, e testes de aderência, conforme orienta o próprio Instrumento *Proknow-C*.

Todavia, verificou-se que o Eixo 3, do modo como estava, acabava retornando trabalhos muito variados, o que tornava as primeiras etapas de filtragem, um pouco mais complicadas. Assim, procedeu-se à separação do Eixo 3 em Eixo 3-A, tratando apenas de Problemas/Eventos, e o Eixo 3-B tratando de Processos/Operações – observar Quadro 2 – tendo os demais eixos, sendo mantidos, da forma como constavam inicialmente.

Quadro 3 - Eixos 3-A e 3-B

3-A	3-B
Eventos/Problemas	**Processos/Operações**
"disaster" OR "emergenc" OR "crisis" OR "problem*"OR "incident*"*	*operation* OR process* OR management OR service**

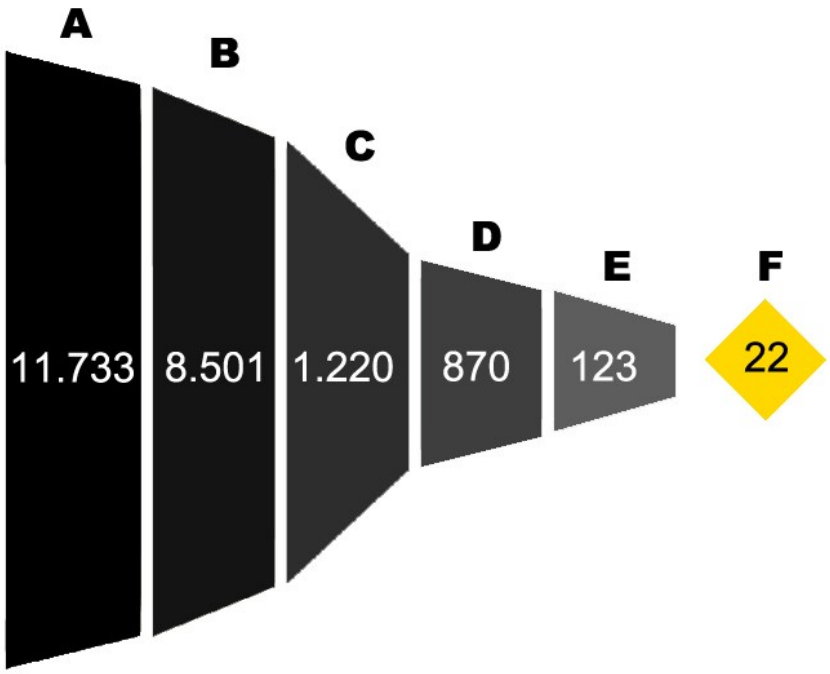
Fonte: dados de pesquisa (2021)

Na prática, o que ocorreu, foi a rodagem de 2 pesquisas “paralelas”, uma, considerando o Eixo 3-A e demais eixos, e a outra, considerando o Eixo 3-B, e demais eixos. Tal processo facilitou a filtragem dos trabalhos, pela leitura dos títulos, pois acabou tornando os retornos das pesquisas nas bases mais homogêneo.

Posterior à filtragem pela leitura dos títulos, os artigos restantes foram reunidos novamente, e o processo seguiu de maneira mais canônica até o fim.

As etapas de filtragem que se seguiram – até os trabalhos selecionados para composição do Portfólio Bibliográfico – são ilustradas, em termos de macro-etapas, conforme consta na Figura 1, abaixo:

Figura 1 - Afunilamento e Filtragem *Proknow-C*



Fonte: dados de pesquisa (2021)

Cada letra indica, uma Macro-Etapa do processo de filtragem, sendo:

A Portfólio Bruto;

B Portfólio Bruto sem duplicidades;

C Prévia Portfólio após leitura dos Títulos;

D Portfólio após leitura dos Títulos;

E Portfólio após leitura dos Resumos;

F Portfólio Final (após leitura integral)

Tais denominações foram uma forma mais gráfica, de ilustrar o processo de seleção dos artigos que compõe o PB.

Da Figura 1, já se pode perceber que, ao fim dos procedimentos de filtragem, os trabalhos selecionados somaram um total de 22 artigos científicos de periódicos revisados.

Da etapa A, optou-se por, na filtragem, dentro das bases de dados, somente considerar artigos de periódicos – e não capítulos de livro e artigos de congressos – por haver o entendimento de que tais trabalhos possuem um peso científico mais preponderante, o que agrega na qualidade do portfólio final e acaba tornando o trabalho de seleção mais célere.

As pesquisas dentro das bases de dados consideraram o operador booleano *AND*, entre os Eixos, e *OR*, intraeixo, ou seja, entre as palavras-chave de cada eixo. Os resultados foram exportados para a ferramenta de gerenciamento bibliográfico *Mendeley*, em formato *.bibtex*.

Foi verificado, inicialmente, um total de 3.963 artigos, com o Eixo 3-A, e 7.770 artigos com o Eixo 3-B, quantidades que, somadas, resultam no quantitativo da etapa A da Figura 1. Fazendo a exclusão das duplicidades, restaram, na pesquisa feita com o Eixo 3-A, 2.716 artigos, e, na pesquisa feita com o Eixo 3-B, 5.785 artigos com o Eixo 3-B, quantidades estas que, somadas, resultam no total de 8.501 artigos não duplicados, conforme discriminado, também, na etapa B da Figura 1.

Procedeu-se, então à filtragem dos títulos, descartando, desde já, aqueles considerados não alinhados seja por se referirem à área de conhecimento absolutamente diversa (neurofisiologia, por exemplo), ou não mencionar nenhum dos eixos temáticos. Como já comentado, tal procedimento foi feito separadamente, ou seja, filtrando por título considerando o Eixo 3-A, e, posteriormente, filtrando por título considerando o Eixo 3-B.

Ao final da filtragem por títulos, restaram, 515 e 715 artigos, considerando as pesquisas feitas com os Eixos 3-A e 3-B, respectivamente, que, somados, resultam em 1.220 artigos, conforme demonstrado na etapa C, da Figura 1.

Entretanto, ressalta-se o quanto já dito, que, neste momento, foram unidos ambos os resultados num conjunto só, quando então, foi verificado que, dentro destes 1220 artigos, havia 350 duplicidades. Estas duplicidades, então, foram excluídas, restando, para a sequência de procedimentos, 870 artigos, conforme ilustrado na etapa D, da Figura 1.

Destes 870 artigos, foi buscado, no *google scholar*, a quantidade de citações de cada um, quando, então, ordenados de modo decrescente, fora calculada a frequência acumulada do conjunto e estabelecida uma representatividade de 92,64%, por meio da qual 374 artigos passaram a compor o Repositório K, e 496 passaram a compor o Repositório P.

Acerca destes repositórios, o primeiro é composto pelo conjunto de artigos que concentram 92,64% – representatividade selecionada – de todas as citações do conjunto, ou seja, é o repositório com validade científica comprovada, do qual serão lidos todos os resumos. Já o segundo, por analogia, é composto pelos trabalhos sem validade científica comprovada, do qual somente serão lidos os artigos mais recentes e os artigos cuja autoria for de algum dos autores do Banco de Autores.

Fora estabelecido, então, que, do Repositório P, seriam lidos os resumos de todos os artigos publicados nos anos de 2020 e 2021, os quais, se acredita, podem não ter cumprido o *threshold* da representatividade em função do pouco tempo em que estão disponíveis. Nesta condição, 212 artigos foram encontrados.

O Banco de Autores é formado pelos autores dos artigos do repositório K, que, pela leitura do resumo, estão alinhados com o que se pretende pesquisar. Assim, é feito um cruzamento entre os autores dos artigos “antigos”, do repositório P (publicados no de 2019 e anteriores), e os autores do Banco de Autores. Deste processo, foram identificados 9 artigos, dos quais apenas 2 demonstraram alinhamento.

Deste modo, tendo feito as etapas acima, foram lidos os resumos de 595 (quinhentos e noventa e cinco) trabalhos, os quais 374 do Repositório K + (mais) 212 do Repositório P “Recentes” + 9 do Repositório P “Autores do Banco de Autores”.

Tendo sido feita a leitura dos resumos, foram descartados 472 trabalhos, restando, para leitura da íntegra do trabalho, 123 artigos, conforme explicitado na Etapa E, da Figura 1.

Passando à última etapa de filtragem, foram tomados 123 artigos na íntegra, dentre os quais, restaram 22 para composição do portfólio bibliográfico, conforme última etapa da Figura 1 apresenta (etapa F). Acredita-se que a considerável redução no número de artigos seja devida, em grande medida, a dois critérios de exclusão adotados, que foram: a exclusão de trabalhos que não tinham enfoque em dados de Redes Sociais (artigos que abrangiam, em suas proposições, dispositivos da *Internet of Things* (IoT) e outras fontes da *Big Data*, por exemplo, foram excluídos), assim

como a exclusão de trabalhos que não propunham abordagens mais empíricas, como *frameworks*, modelos ou processos voltados ao tratamento dos dados advindos de Redes Sociais.

A lista dos artigos selecionados, que compõe o PB sãos os discriminados no Quadro 3, logo abaixo:

Quadro 4 - Artigos Selecionados *Proknow-C*

Artigo	Autoria e Ano
A framework to simplify pre-processing location-based social media big data for sustainable urban planning and management	(ABDUL-RAHMAN et al., 2021)
A sentiment reporting framework for major city events: Case study on e China-United States trade war	(ALKHATIB et al., 2020)
A volunteered geographic information framework to enable bottom-up disaster management platforms	(POORAZIZI; HUNTER; STEINIGER, 2015)
An Arabic social media based framework for incidents and events monitoring in smart cities	(ALKHATIB; EL BARACHI; SHAALAN, 2019)
Analysis of social networking service data for smart urban planning	(MORA et al., 2018)
Assessing the Impact of Transportation Diversity on Postdisaster Intraurban Mobility	(RAHIMI-GOLKHANDAN; GARVIN; WANG, 2021)
Can we predict a riot? Disruptive event detection using twitter	(ALSAEDI; BURNAP; RANA, 2017)
Creating corroborated crisis reports from social media data through formal concept analysis	(ANDREWS et al., 2016)
CrowdPulse: A framework for real-time semantic analysis of social streams	(MUSTO et al., 2015)
Detection of barriers to mobility in the smart city using twitter	(SÁNCHEZ-ÁVILA et al., 2020)
Digital Earth from vision to practice: Making sense of citizen-generated content	(CRAGLIA; OSTERMANN; SPINSANTI, 2012)
Disaster-resilient communication ecosystem in an inclusive society – A case of foreigners in Japan	(SAKURAI; ADU-GYAMFI, 2020)
Effective surveillance and predictive mapping of mosquito-borne diseases using social media	(JAIN; KUMAR, 2018)
Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data	(HASNAT; HASAN, 2018)
Internet of people enabled framework for evaluating performance loss and resilience of urban critical infrastructures	(YUAN et al., 2021)
Social Network Based Crowd Sensing for Intelligent Transportation and Climate Applications	(TSE et al., 2018)
Social Sensing in Disaster City Digital Twin: Integrated Textual–Visual–Geo Framework for Situational Awareness during Built Environment Disruptions	(FAN; JIANG; MOSTAFAVI, 2020)
Tracking Flooding Phase Transitions and Establishing a Passive Hotline with AI-Enabled Social Media Data	(WANG et al., 2020)
TwitterSensing: An Event-Based Approach for Wireless Sensor Networks Optimization Exploiting Social Media in Smart City Applications	(COSTA et al., 2018)
Using adverse weather data in social media to assist with city-level traffic situation awareness and alerting	(LU et al., 2018)
Using digital footprints for a city-scale traffic simulation	(MCARDLE et al., 2014)
Using social media for emergency response and urban sustainability: A case study of the 2012 Beijing rainstorm	(WANG et al., 2016)

Fonte: dados de pesquisa (2021)

2.1.2 Modelagem de Tópicos para Revisão de Literatura

A segunda estratégia para revisão de literatura é a Revisão de Literatura que faz uso da modelagem de tópicos para identificar *clusters* de documentos cujos assuntos abordados são similares. A modelagem de tópicos foi feita utilizando o método de *machine learning* não supervisionado denominado *Latent Dirichlet Alocation* (LDA) (BLEI; NG; JORDAN, 2003) para análise de tópicos de um corpus de documentos. Este *corpus* de documentos são, na prática, *abstracts* de artigos científicos, os quais foram selecionados de acordo com procedimentos que serão detalhados na sequência.

A presente abordagem para RSL segue a *framework* para revisão de literatura – considerando modelagem de tópicos – proposta por Asmussen e Möller (2019), baseada em técnica de mineração de texto, cujas etapas compreendem – tendo um portfólio bruto de artigos selecionados – as etapas de pré-processamento e modelagem de tópicos. Os passos são descritos na sequência, porém, antes, é necessário delinear a estratégia até conseguir chegar ao portfólio bruto de artigos.

Para realização da presente abordagem de RSL, o primeiro passo é a escolha da base de dados para busca de artigos. Neste caso, já pensando na posterior análise via Modelamento de Tópicos, e considerando aspectos de formatação, foi escolhida, apenas, a base de dados *Scopus*. A razão para tanto continua sendo o fato de se tratar de base de dados com amplo número de artigos, bastante multidisciplinar.

Tendo sido feitas a escolha da base de dados, passa-se, então, para a definição de eixos e palavras-chave. Inicialmente, foram tomados os mesmos eixos e palavras-chave constantes no Quadro 1, contudo, tendo em vista ter sido estabelecido um *threshold*/teto de resultados, de no máximo 1000 artigos para aplicação da modelagem de tópicos, algumas alterações se fizeram necessárias. A delimitação de um teto máximo de resultados para que fosse rodada modelagem de tópicos é devida tendo por parâmetro o próprio trabalho de Asmussen e Möller (2019), onde os autores, rodaram a LDA para 650 artigos, com tempo de processamento de três horas e meia, o que já é considerado um tempo razoável, na medida em que este poderia ser dedicado à leitura e filtragem dos títulos dos artigos, caso fosse se seguir uma abordagem manual. Assim, uma quantidade muito grande de artigos para então rodar

a LDA, poderia acabar desvirtuando a comentada celeridade desta revisão por modelagem de tópicos.

As sequências de busca e seus resultados, na base *Scopus*, culminaram, ao fim nos eixos e palavras-chave constantes no Quadro 4, abaixo, onde se nota – em virtude do teto de resultados acima comentado – a redução do número de palavras-chave por eixo, em comparação à revisão de literatura pelo *Proknow*-C:

Quadro 5 - Eixos e Palavras-chave para coleta na base *Scopus* e LDA

EIXOS			
1	**2**	**3**	**4**
Dados de Redes Sociais; Human/Social Sensing	**Abrangência Cidades**	**Eventos Problemas**	**Inteligência**
"social media" OR "sns" OR "human sens" OR "social sens*"*	*"cit*" OR "region*" OR "municip*" OR "govern*" OR "local*" OR "urban*" OR "communit*"*	*"disaster" OR "emergenc*" OR "incident*" OR process* OR management*	*"smart*" OR "sustainab*"*

Fonte: dados de pesquisa (2021)

A pesquisa obteve 1481 resultados, considerando as palavras chaves acima elencadas. Foram mantidos, em todas as buscas, os mesmos eixos, sendo que a mudança ocorria intra-eixo, pela manutenção de apenas parte de todas as *Keywords* inicialmente estipuladas para cada eixo.

Tendo em vista, ainda, o resultado acima ultrapassar o limite de 1000 artigos, pré-estabelecido, foi utilizado o critério de considerar apenas artigos de *Journals*, escolha que também foi adotada na revisão pela sistemática do *Proknow-C* e que, por entender que estes trabalhos possuem maior relevância científica, agregando mais à qualidade do Portfólio, não comprometendo a qualidade da seleção, portanto, a desconsideração de trabalhos de congressos e capítulos de livro.

Deste modo, o quantitativo final de artigos levantados, nesta primeira etapa, foi equivalente a 674 trabalhos, os quais compuseram o conjunto de documentos sob os quais seria performada a modelagem de tópicos.

Partindo da *framework* de Asmussen e Möller (2019), o que se fez, então, foi aplicar o método de *Machine Learning* não supervisionado denominado *Latent Dirichlet Alocation* (LDA) – procedimento de mineração de textos – para realizar a "clusterização" do *corpus*, sem antes, contudo realizar a etapa prévia, de pré-processamento.

Nesta etapa, buscou-se limpar o conjunto de *abstracts* de modo a deixar os dados de *input* à análise mais homogêneos. As ações de pré-processamento de dados que foram performadas compreendem a normalização de todo o texto em minúsculo, remoção de *urls*, lematização (redução de cada palavra a seu "lema"/assunto principal), tokenização, remoção de *stop words* (preposições e artigos, por exemplo) e remoção de pontuação.

Optou-se também, pela inclusão das palavras-chave dos eixos de pesquisa na base de dados, na lista de *stop words*, por entender que estes termos seriam constantes em todos os tópicos posteriormente identificados.

Assim, os termos mais frequentes do *dataset* de *abstracts* puderam ser visualizados por meio de uma *word cloud*, a qual permitia a visualização de outros termos que porventura fossem reputados enquanto não significativos, os quais seriam, por sua vez, acrescidos à lista de *stop words*. O referido processo de pré-processamento, portanto, é iterativo.

A *word-cloud* final, considerando a iteração nesta etapa de pré-processamento pode ser vista abaixo, na Figura 2:

Figura 2 - Nuvem de Palavras *Topic Modelling Literature Review*

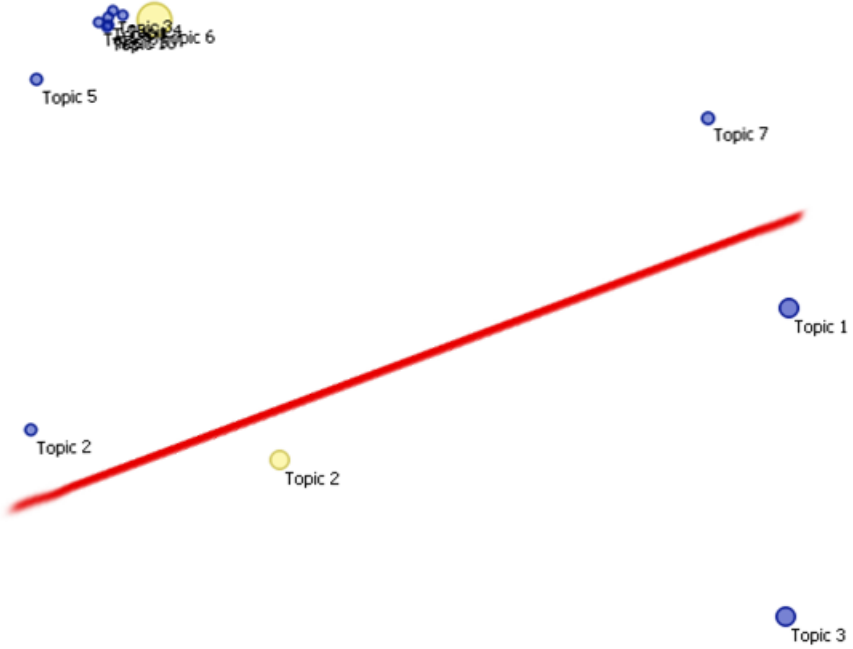
Fonte: dados de pesquisa (2021)

Tendo sido feito o pré-processamento do conjunto de *abstracts*, passou-se então à realização de modelagem de tópicos por meio da LDA.

Os procedimentos para aplicação da LDA são mais detalhadamente definidos no capítulo 3, da Metodologia. De modo geral, a estratégia aqui adotada, foi a definição, inicial, de um número K de tópicos a serem extraídos do *dataset.* Tendo

sido definidos 10 tópicos, inicialmente, observou-se a distribuição abaixo demonstrada, acima do traço vermelho, por meio do qual foi possível a verificação da existência, na prática, de três grupos distintos de tópicos, a julgar pela distância e agrupamento verificados entre os tópicos.

Figura 3 - Distribuição de Tópicos MDS *Topic Modelling Literature Review*



Fonte: dados de pesquisa (2021)

Levando em consideração que o *set* de K=10 tópicos retornou uma distribuição que se distanciou entre si, em três "cantos" distintos (Figura 3), a definição de K=3 se tornou lógica, e foi exatamente o que foi feito, posteriormente, gerando a distribuição observada que consta abaixo da linha vermelha, conforme Figura 3.

A visualização da distribuição gráfica dos tópicos é feita por meio da *Multidimensional Scaling* (MDS) (WICKELMAIER, 2003), conforme também tratado no capítulo 3 da Metodologia. A visualização com o MDS não é uma estratégia definitiva, nem inequívoca, mas permite, até mesmo ao usuário mais leigo, a análise gráfica da distribuição tópica, sendo adequada para uma razoável assertividade da delimitação de K tópicos.

Tendo sido definidos 3 tópicos, estes devem ser rotulados. Para executar tal passo, passou-se a analisar as palavras mais significativas/representativas de cada tópico, bem como os títulos dos artigos com maior *score* de pertencimento, para cada tópico, e de seus *abstracts* correspondentes.

Os tópicos e seus rótulos são os que constam no Quadro 5, abaixo:

Quadro 6 - Tópicos *Topic Modelling Literature Review*

TÓPICO	RÓTULO	NÚMERO DE ARTIGOS
TÓPICO 1	*Digital Government, Gov 3.0 e Technology Acceptance*	379
TÓPICO 2	*Social Media data analysis, City environment, Incidents and Sentiments*	647
TÓPICO 3	*Environmentalism and Social Responsability*	322

Fonte: dados de pesquisa (2021)

Tendo em vista os propósitos de pesquisa desta dissertação, definiu-se desde logo, o Tópico 2 como sendo o tópico de interesse, segundo o qual os próximos critérios de seleção do portfólio bibliográfico seriam aplicados. Verificou-se, desde logo, referente aos tópicos, que todos os 647 trabalhos tinham algum nível de alinhamento, por menor que fosse, com o Tópico 2, relação que não se fez presente, referente ao alinhamento dos artigos aos Tópicos 1 e 2.

Junto da seleção do Tópico 2, foi determinado um *threshold* mínimo de alinhamento, do artigo com o tópico, no valor de 0,8, ou seja, artigos que não tivessem um *score* de pertencimento de no mínimo 80% com o Tópico 2 foram sumariamente desconsiderados. Com isso, o número de artigos caiu de 674 para 381, com este filtro.

Ainda, foi definido um critério de atualidade, para seleção dos artigos, de modo a eliminar artigos antigos. O *threshold,* desta vez, tratou de considerar apenas artigos publicados no ano de 2016 ou posteriores, excluindo os que não cumpriam este critério. O número de artigos passou de 381, para 335, com este filtro.

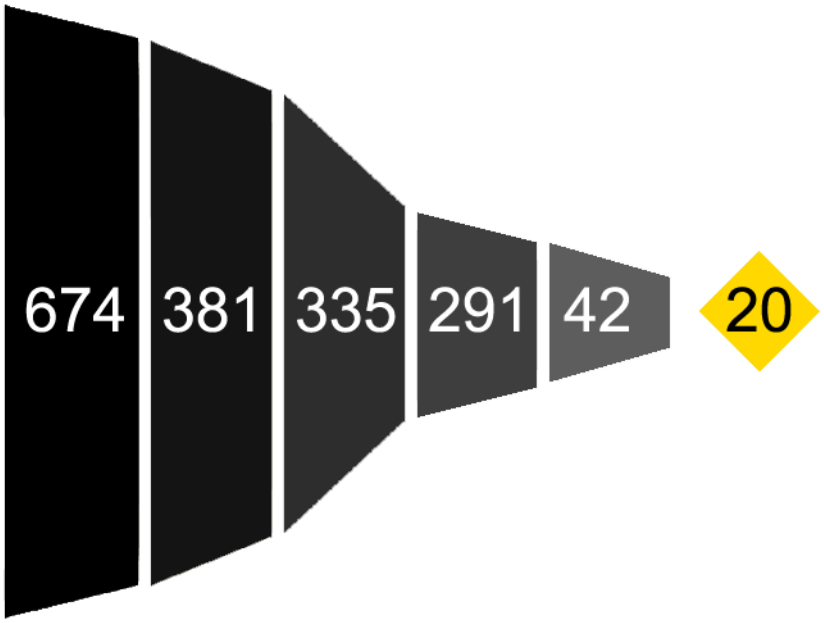
O próximo critério foi o de representatividade científica, definida, neste caso, em função do número de citações dos artigos que ainda compunham o portfólio bruto. Foi definida uma representatividade de 95,6%, o que significa dizer que dos 335 artigos restantes, 184 deles detinham 95,6% do somatório de citações do portfólio, sendo que, os 151 restantes, foram separados em um repositório próprio, chamado aqui de Repositório 2.

No Repositório 2 constaram, então, nesta etapa, os artigos sem representatividade científica validada, contudo, necessário levar em consideração o ano de publicação destes. Estabeleceu-se que artigos recentes – entendidos, aqui, enquanto aqueles publicados de 2020 em diante – contidos no Repositório 2, possivelmente não possuíam número de citações que comprovasse sua validade científica, em função de serem muito recentes, razão pela qual foram resgatados.

Assim, dos 151 artigos do Repositório 2, 107 foram salvos, considerando critério de serem recentes, e os 44 restantes, de anos anteriores a 2020, foram descartados. Restaram, então, para leitura dos *abstracts*, 291 artigos.

Tendo sido feita a leitura dos resumos, foram descartados outros 249 artigos, restando, para leitura completa, 42 trabalhos. Destes 42, foram selecionados, ao final, 20 artigos. O afunilamento deste segundo processo de revisão sistemática pode ser visualizado, em termos de macro-etapas, conforme consta na Figura 4, abaixo:

Figura 4 - Afunilamento e Filtragem do Tópico 2



Fonte: dados de pesquisa (2021)

A lista final de artigos, depois de performadas todas as etapas de filtragem, pode ser conferida no Quadro 6.

Quadro 7 - Artigos Selecionados usando Modelagem de Tópicos – LDA: tópico 2

Artigo	Autoria e Ano
A framework to simplify pre-processing location-based social media big data for sustainable urban planning and management	(ABDUL-RAHMAN et al., 2021)
A sentiment reporting framework for major city events: Case study on the China-United States trade war	(ALKHATIB et al., 2020)
An Arabic social media based framework for incidents and events monitoring in smart cities	(ALKHATIB; EL BARACHI; SHAALAN, 2019)
Beyond sound level monitoring: Exploitation of social media to gather citizens subjective response to noise	(GASCO et al., 2019)
Capturing citizen voice online: Enabling smart participatory local government	(ALIZADEH; SARLAR; BIRGPUME, 2019)
Creating corroborated crisis reports from social media data through formal concept analysis	(ANDREWS et al., 2016)
CrowdPulse: A framework for real-time semantic analysis of social streams	(MUSTO et al., 2015)

Detection of barriers to mobility in the smart city using twitter	(SÁNCHEZ-ÁVILA et al., 2020)
Extraction method and integration framework for perception features of public opinion in transportation	(LIU; TENG; GONG, 2021)
Framing twitter public sentiment on Nigerian government COVID-19 palliatives distribution using machine learning	(ADAMU et al., 2021)
From sustainable development goals to sustainable cities: A social media analysis for policy-making decision	(MARZOUKI et al., 2021)
Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data	(HASNAT; HASAN, 2018)
Internet of people enabled framework for evaluating performance loss and resilience of urban critical infrastructures	(YUAN et al., 2021)
Linking social, semantic and sentiment analyses to support modeling transit customers' satisfaction: Towards formal study of opinion dynamics	(EL-DIRABY; SHALABY; HOSSEINI, 2019)
Review of discussions on internet of things (IoT): Insights from twitter analytics	(JOSEPH et al., 2017)
The Missing Parts from Social Media–Enabled Smart Cities: Who, Where, When, and What?	(YUAN et al.,2020)
TwitterSensing: An Event-Based Approach for Wireless Sensor Networks Optimization Exploiting Social Media in Smart City Applications	(COSTA et al., 2018)
Use of social media for assessing sustainable urban mobility indicators	(SDOUKOPOULOS et al., 2018)
Using social media for emergency response and urban sustainability: A case study of the 2012 Beijing rainstorm	(WANG et al., 2016)
What causes different sentiment classification on social network services? Evidence from weibo with genetically modified food in China	(LI et al., 2020)

Fonte: dados de pesquisa (2021)

Retornaram em ambas as revisões – tanto por meio do *Proknow*-C, quanto da Modelagem de Tópicos (LDA) – oito artigos iguais.

***

A presente dissertação se utiliza de dois grandes grupos teóricos, que são: *Smart Sustainable Cities* e Redes Sociais, sendo que o primeiro apresenta conceitos relacionados à gestão pública urbana e o segundo trata, a nível macro, da fonte de dados utilizada nesta pesquisa, explicando sua relevância enquanto elemento de TIC, com potencial aplicação nas cidades.

Na sequência, este capítulo de revisão de literatura se desdobra em mais três partes, sendo as duas primeiras para tratar dos dois grupos citados, e a terceira onde serão abordados trabalhos que fizeram uso de dados de redes sociais, num contexto de cidades.

2.2 Cidades Inteligentes e Sustentáveis

Tem-se certo que o crescimento populacional urbano, que vem sendo uma constante há décadas, acentuar-se-á ainda mais no futuro, havendo a perspectiva de que, quando for atingido o ano de 2050, a população urbana global, que hoje corresponde 55 (cinquenta e cinco) por cento, a nível mundial, chegue na marca de 68 (sessenta e oito) por cento (UN, 2018).

Na esteira do crescimento urbano, crescem também, de igual modo, os desafios associados à gestão das cidades, o que tem feito com que, as cidades tenham passado a adotar conceitos de cidades inteligentes, para solucionar estes novos desafios (MENDONÇA et al., 2016), sendo aquelas, definidas – conceitualmente – enquanto “comunidades de razoável tecnologia, interconectadas, sustentáveis, confortáveis, atrativas e seguras” (LAZAROIU; ROSCIA, 2012, p.326).

Numa definição basilar acerca de *smart cities*, é trazida uma abordagem que abarca 6 (seis) dimensões de cidades inteligentes, que seriam economia inteligente, pessoas inteligentes, governança inteligente, mobilidade inteligente, meio-ambiente inteligente e vida inteligente, segundo a qual, uma cidade que é, de fato inteligente, deve ter bom desempenho nestas seis óticas, sem deixar, ainda, de buscar o constante melhoramento em cada uma delas (GIFFINGER et al., 2007).

Ainda, é defendida a ideia de que a cidade inteligente é aquela em que a “tecnologia está a serviço das pessoas e de sua melhoria na qualidade de vida, econômica e social” (LAZAROIU; ROSCIA, 2012, p.332).

Na mesma linha, Batty et al. (2012) defendem que numa *smart city* há a mescla entre as infraestruturas físicas tradicionais das cidades, com as Tecnologias de Informação e Comunicação (TIC), de maneira integrada e coordenada, fazendo uso das novas tecnologias digitais.

Caragliu, Del Bo e Nijkamp (2009) complementam o raciocínio acima, afirmando que o desempenho urbano não depende somente da existência da infraestrutura física, mas também e cada vez mais de uma estrutura lógica e social, de comunicação e conhecimento. Indo mais além, uma cidade inteligente investe tanto em capital humano e social, como na modernização das infraestruturas físicas e de TIC, de modo a prover a qualidade de vida, com adequada gestão dos recursos naturais, prezando pelo crescimento sustentável, e utilizando-se, também, da governança participativa (CARAGLIU; DEL BO; NIJKAMP, 2009).

Batty et al., (2012) agregam à discussão, argumentando em função de se levar em conta novas fontes de dados urbanos, e que sejam articulados aos problemas,

políticas e planos para a cidade, tendo ainda a preocupação com o engajamento com a comunidade, sabendo, também, que este requer novas formas de participação online, na qual sejam utilizadas as novas tecnologias de TIC, aspecto este que intersecta não só os propósitos do presente trabalho, como também a dimensão de governança inteligente.

Acerca da dimensão *Smart Governance,* entende-se que esta compreende "aspectos de participação política, de seviços aos cidadãos e do funcionamento da administração" (GIFFINGER et al., 2007, p.11). Abstrai-se, deste modo, fazendo a delimitação dos dados advindos de Redes Sociais enquanto subconjunto destas novas tecnologias de TIC e fontes alternativas de informação das cidades, que o propósito deste trabalho intersecta a dimensão de governança inteligente, referente à *smart cities*.

2.3 Redes Sociais

Como já comentado, um subgrupo do grupo de ferramentas tecnológicas da TIC é composto pelas Redes Sociais digitais, que mesmo possuindo diferenças entre si, usualmente, permitem aos usuários "1) construir um perfil público ou semipúblico [...], 2) articular uma lista composta de outros usuários com os quais eles compartilham uma conexão e 3) visualizar e percorrer sua lista de conexões e aquelas feitas por outras pessoas, dentro do sistema" (BOYD; ELLISON, 2007, p.211).

É dito também que o termo Redes Sociais é um conceito "guarda-chuva", na medida que engloba uma grande variedade de diferentes plataformas, as quais ganharam uma notória atenção na última década, permitindo aos seus usuários a conexão mútua (ALIZADEH; SARLAR; BIRGPUME, 2019).

Outra faceta das redes sociais diz respeito, também, a natureza dinâmica que possuem, segundo o qual, sob uma ótica organizacional, possibilitam o envolvimento tanto de cima para baixo como de baixo para cima (ALIZADEH; SARLAR; BIRGPUME, 2019).

Os dados produzidos pelos usuários de Redes Sociais têm por característica refletir, na maior parte das vezes, o dia a dia das pessoas (MORA et al., 2018), e serem bastante variados, desde conteúdo textual, audiovisual até metadados, também são produzidos em alta velocidade, de maneira ininterrupta, além de serem

produzidos em grande volume, o que remonta os aos 3 (três) “Vs” característicos da Big Data (WANG et al., 2016).

A produção de dados nas Redes sociais conta com ampla difusão e disseminação (LU et al., 2018, KANKANAMGE et al., 2020), sendo ao mesmo tempo um desafio e uma oportunidade, para diversas organizações (tanto públicas quanto privadas) “fazer sentido” destes dados, de modo a serem úteis à tomada de decisão (ALSAEDI; BURNAP; RANA, 2017).

Um fator preponderante no aumento do volume de dados de redes sociais, foi a maior disponibilização dos *smartphones*, permitindo que redes sociais como *Twitter* e *Facebook* se tornassem ainda mais populares, abarcando bilhões de usuários ao redor do mundo, os quais, por sua vez, utilizam daquelas para expor suas opiniões ou experiências, o que, do ponto de vista organizacional, dá meios de se utilizar tais informações para melhoramento dos produtos e serviços a serem entregues (MORA et al., 2018).

Não só do ponto de vista prático, mas também do ponto de vista acadêmico, trabalhos que se utilizem de dados de Redes Sociais têm sido feitos, aproveitando-se da “onda” de conteúdos gerados pelos próprios usuários, como conteúdos textuais, o que fez com que pesquisadores de uma miríade de campos do conhecimento tratassem de propor diferentes análises que vão desde tratar o comportamento humano, opinião pública até *marketing* e campanhas políticas (SDOUKOPOULOS et al., 2018).

Observa-se, ainda, na literatura tomada, que a maior parte dos trabalhos fazem uso de dados advindos da Rede Social *Twitter* (ALKHATIB et al., 2020, POORAZIZI; HUNTER; STEINIGER, 2015, ALKHATIB; EL BARACHI; SHAALAN, 2019, ALSAEDI; BURNAP; RANA, 2017, ANDREWS et al., 2016, SÁNCHEZ-ÁVILA et al., 2020, CRAGLIA; OSTERMANN; SPINSANTI, 2012, JAIN; KUMAR, 2018, FAN; JIANG; MOSTAFAVI, 2020, WANG et al., 2020, COSTA et al., 2018 LU et al., 2018, WANG et al., 2016, YUAN et al., 2021, RAHIMI-GOLKHANDAN; GARVIN; WANG, 2021, SAKURAI; ADU-GYAMFI, 2020, HASNAT; HASAN, 2018, TSE et al., 2018, MCARDLE et al., 2014, MUSTO et al., 2015, ABDUL-RAHMAN et al., 2021, JOSEPH et al., 2017, EL-DIRABY; SHALABY; HOSSEINI, 2019, YUAN et al.,2020, ALIZADEH; SARLAR; BIRGPUME, 2019, GASCO et al., 2019, SDOUKOPOULOS et al., 2018, LI et al., 2020, LIU; TENG; GONG, 2021, ADAMU et al., 2021, MARZOUKI et al., 2021).

Acredita-se isso ser devido, muito em função do grande volume de dados produzidos diariamente por seus inúmeros usuários (ALSAEDI; BURNAP; RANA, 2017, HASNAT; HASAN, 2018, MORA et al., 2018), assim como, também, em razão de seu aspecto mais *open source*, se comparada ao *Facebook*, por exemplo, uma vez que o *Twitter* disponibiliza uma *Application Programming Interface* (API) própria, de fácil utilização, que permite aos interessados fazerem requisições ao *back-end*, com razoável facilidade (KANKANAMGE et al., 2020).

No subitem seguinte, passa-se a uma breve revisão sobre trabalhos que se utilizaram de dados de Redes Sociais, e aplicaram num contexto de cidades.

2.4 Dados de Redes Sociais no Contexto das Cidades

Conforme já comentado no capítulo 1 deste trabalho, quando se está levando em conta o contexto das cidades, a literatura tem abordado a temática das redes sociais de diferentes formas. Há trabalhos com enfoque na proposição de soluções para eventos que ocorrem nas cidades (POORAZIZI; HUNTER; STEINIGER, 2015, ALKHATIB; EL BARACHI; SHAALAN, 2019, ALSAEDI; BURNAP; RANA, 2017, ANDREWS et al., 2016, SÁNCHEZ-ÁVILA et al., 2020, CRAGLIA; OSTERMANN; SPINSANTI, 2012, JAIN; KUMAR, 2018, FAN; JIANG; MOSTAFAVI, 2020, WANG et al., 2020), como desastres, emergências e incidentes de um modo geral, assim como são, também, encontrados trabalhos com foco mais diversificado, propondo novos processos que façam uso de dados de Redes Sociais, para a consecução de objetivos (HASNAT; HASAN, 2018, MORA et al., 2018, MUSTO et al., 2015, ABDUL-RAHMAN et al., 2021), tendo ainda aqueles que optem por um enfoque maior à opinião e sentimento públicos (ALKHATIB et al., 2020, MUSTO et al., 2015, ABDUL-RAHMAN et al., 2021, JOSEPH et al., 2017, EL-DIRABY; SHALABY; HOSSEINI, 2019, ALIZADEH; SARLAR; BIRGPUME, 2019, SDOUKOPOULOS et al., 2018, LI et al., 2020, LIU; TENG; GONG, 2021, ADAMU et al., 2021).

2.4.1 Cidades Inteligentes e Dados

Se comenta que um desafio com relação às cidades inteligentes e sua necessidade de conseguir obter “inteligência”, se refere à habilidade de conseguir acessar informações, tempestivamente, da população, eventos, ativos e outros, de

maneira eficiente e escalável (ALKHATIB et al., 2020). Neste interim, se observa que as maneiras mais usuais de obtenção de dados – como *surveys* – acabam apresentando desvantagens, na medida em que consomem mais recursos financeiros para elaboração, aplicação e análise, tomam mais tempo e possuem um viés praticamente inevitável nas questões (ALKHATIB et al., 2020).

Por outro lado, se verifica o contexto das cidades enquanto um celeiro bastante pujante de produção de informações, na medida em que os cidadãos, que estão vivendo suas vidas diárias, acabam compartilhando suas opiniões e sentimentos, em Redes Sociais, durante eventos em curso e outras situações e, os quais se tornam informações úteis que devem ser levadas em conta pelos planejadores das cidades (MUSTO et al., 2015, COSTA et al., 2018, ALKHATIB et al., 2020).

A verificada imprescindibilidade na utilização da inteligência advinda das pessoas, por meio de dados de Redes Sociais, torna-se ainda mais preponderante quando se está diante dos pressupostos de Cidades Inteligentes (e sustentáveis). Nestas, as decisões devem levar em conta o coletivo, tendo os cidadãos um papel cada vez mais central, a ponto de o sucesso de projetos nas cidades, depender de haver inequívoco engajamento e análise das necessidades da comunidade (EL-DIRABY; SHALABY; HOSSEINI, 2019).

Acredita-se que os dados de Redes Sociais, enquanto potencial fonte de dados contendo os anseios e necessidades da população, fornecem meio para que o poder público tome decisões mais acertadas, guiando o processo de planejamento de ações, refletindo ainda os requisitos de sustentabilidade, que são próprios da utilização desta fonte de dados para obter informação útil ao processo decisório (EL-DIRABY; SHALABY; HOSSEINI, 2019; ADAMU et al., 2021).

2.4.2 Dados Produzidos Em Redes Sociais

Tomando o ganho da já mencionada facilidade, velocidade e baixo custo trazidos pela utilização dos dados de Redes Sociais enquanto fonte, frente à métodos tradicionais como *surveys* (GASCO et al.,2019), verifica-se que as características inerentes àquela fonte de dados fornece potencial a aplicações em contextos que necessitam de uma avaliação sobre uma situação momentânea e resposta rápida,

como em eventos do tipo desastre, por órgãos de atendimento à emergência (WANG et al., 2016, ANDREWS et al., 2016).

Indo na mesma linha, Yuan et al., (2021), abordando ainda o conceito de Cidade Resiliente, afirmam ser necessário à uma cidade assim classificada, de modo a ser capaz de responder a desastres, ter sistemas de resposta que façam uso intensivo de dados que lhes permitam a identificação tempestiva de incidentes, intersectando assim, as características próprias dos dados de Redes Sociais.

Os dados de Redes Sociais têm por característica intersectarem os conceitos de *Big Data*, no sentido de que são produzidos de maneira veloz, em grande volume e com tipos variados (MUSTO et al., 2015), e, tendo em vista estes predicados, sua análise é impraticável sem o auxílio da computação, a qual deve ser empregada para prover informações indicar as relações encontradas, municiando o tomador de decisão (ANDREWS et al., 2016).

Em análises voltadas ao contexto das cidades, uma característica de grande utilidade nos dados de redes sociais está presente – em algumas plataformas, como o *Twitter* – nos metadados das postagens, que diz respeito a localização geográfica das postagens. Trabalhos que façam uso desta característica podem ser bastante úteis, não só à captura das percepções das pessoas de maneira quase instantânea, mas podem fomentar serviços de *Smart City*, havendo um cuidado com os aspectos críticos que envolvem sua utilização, tal qual comentam Yuan et al., (2020) em seu trabalho.

Em se tratando de *Twitter*, deve ser feita a diferenciação do que seria o conteúdo geolocalizado e georreferenciado. O primeiro, entra em consonância com o que é dito no parágrafo acima, ou seja, se trata de informação contendo a exata localização geográfica, com coordenadas, no momento da postagem, pelo usuário, enquanto o segundo, contém informação geográfica no corpo do *tweet* (SÁNCHEZ-ÁVILA et al., 2020). Entretanto, uma limitação observada na literatura, é a de que a minoria dos *tweets* contempla a informação geográfica precisa, em seus metadados, visto este ser um atributo que fica a critério do usuário ativá-lo ou não (SÁNCHEZ-ÁVILA et al., 2020), algo também observado em trabalhos com outras redes sociais similares, como o *Sina Weibo* (WANG et al., 2016).

2.4.3 Líderes de Opinião/Usuários Centrais/Conteúdo de Alto Impacto

Nesse interim, uma abordagem inovadora que tem sido utilizada na literatura, para a obtenção de dados de Redes Sociais, passa pela identificação de usuários centrais ou líderes de opinião, a exemplo do que fizeram Alkhatib et al., (2020). Estes, em seu trabalho, propuseram uma *framework* para monitoramento da opinião pública, tendo por novidade, o foco está nos *posts* de alto impacto de líderes públicos de opinião e seus seguidores, de modo a fornecer entendimento em profundidade e gerar relatórios do sentimento público.

Na mesma linha, há recomendação para que trabalhos futuros façam a identificação destes usuários centrais, ou *influential spreaders*, como dito por Abdul-Rahman et al., (2021), ou grupos de usuários alvo, conforme Yuan et al., (2020), de modo a ter um foco na obtenção das informações, de maneira mais delineada.

A identificação destes usuários (pessoas públicas) e a extração do conteúdo de suas postagens e das interações de outros usuários com estas, pode ser uma alternativa no intento de obter conteúdo georreferenciado.

2.4.4 Opinião Pública

Do ponto de vista prático, e das possibilidades de aplicação advindas da análise de dados de Redes Sociais, se verifica que, por meio da *crowdsourcing of opinions* é possível realizar minerar as opiniões das pessoas e performar análises de sentimento, sobretudo tendo como reduto fonte, o *Twitter*, que é uma rica e volumosa fonte de opiniões dos usuários (ALIZADEH; SARLAR; BIRGPUME, 2019).

Por características, as redes sociais permitem uma fluida e orgânica comunicação entre seus usuários, os quais expressam de maneira livre sua opinião, sobre os tópicos que sejam de seu interesse, e isso fica ainda mais evidente em redes socias do tipo *microblog* – como é o caso do *Twitter* – onde os usuários postam mensagens, limitadamente curtas, sobre o que lhes convém, e, justamente por haver a limitação de caracteres, as postagens tendem a ser mais objetivas e “direto ao ponto” (GASCO et al., 2019).

Neste sentido, a partir da construção de um banco de dados de *tweets*, de modo a extrair conteúdo útil ao tomador de decisão (ALKHATIB et al., 2020), é necessário que seja feito algum tipo de análise do conteúdo destas postagens, de modo a entender a que cada *tweet* se refere (JOSEPH et al., 2017). Para tanto, deve-

se aplicar o Processamento de Linguagem Natural (NLP), para tentar abstrair informação dos dados brutos.

Para além disso, a análise de sentimentos pode ser aplicada individualmente, para cada excerto de texto, de modo a identificar a polaridade de sentimento sobre aquele trecho, se positiva, negativa ou neutra (LI et al., 2020), mas pode, também, ser aplicada em construtos mais significativos como tópicos.

Esta segunda possibilidade remonta ao que definem El-Diraby, Shalaby e Hosseini (2019) enquanto Opinião Pública. Pode-se entender a opinião pública enquanto a intersecção entre um tópico de interesse em um conjunto de dados, e o sentimento da fonte produtora do dado, sobre aquele tópico. Assim, a opinião pública é entendida enquanto “a combinação de uma análise semântica e de sentimento” sobre um dado tópico (EL-DIRABY; SHALABY; HOSSEINI, 2019).

2.4.5 Modelos da Literatura

No portfolio de artigos revisados foram observadas propostas e modelos parecidos com o que aqui será proposto, tendo, cada qual, características próprias. A propria denominação, por vezes chamando de *framework,* e outras chamando de arquitetura ou metodologia, propriamente, mas que, na prática, seguem o mesmo fundamento de descrever etapas que compreendem, geralmente, desde a extração até a análise dos dados de redes sociais.

Trabalhos como os de Costa et al., (2018) e Fan, Jiang e Mostafavi (2020), contemplaram claro foco no auxílio à incidentes. Costa et al., (2018) propuseram uma abordagem para detectar e classificar eventos, tendo por fonte o *Twitter*, e, com essa classificação, prioridades de monitoramento a sensores físicos, os quais podem então monitorar as áreas mais afetadas, de acordo com o *input* dos dados de redes sociais. Já Fan, Jiang e Mostafavi (2020) apresentaram um *framework* para detecção de disrupções – a exemplo do teste do modelo, que foi o caso do furacão *Harvey* em *Houston*, 2017 – nas cidades, com base na frequência de *tweets* sobre dado assunto, combinando, junto da análise textual, a análise das descrições de imagens tuitadas e análise geográfica com plotagem em mapa.

Outro trabalho com abordagem predominantemente em eventos é o de Alkhatib, El Barachi e Shaalan (2019), onde os autores propõem uma *framework* para gestão de eventos e incidentes em cidades inteligentes, *framework* essa que

possibilita o fornecimento de relatórios em tempo real sobre eventos e incidentes – com base em dados do *Twitter* – de modo a prover os órgãos de resposta a emergência com informações úteis para planejamento e tomada de decisão. Um destaque e diferenciação deste trabalho com relação aos dois, acima apresentados, é ter utilizado *feed* textual em árabe, e não em inglês.

Intersectanto tanto a abordagem em eventos como análise de sentimentos, têm-se o trabalho de Musto et al., (2015), no qual é proposta uma *framework* para análise textual, em tempo real, do conteúdo de redes sociais, aplicando-se, também, análise de sentimentos. A *framework* de Musto et al., (2015) foi testada em dois casos: para mapeamento de discursos de ódio, na Itália – com indicação geográfica de onde ocorrem – e a análise do capital social para recuperação do terremoto de Áquila em 2009.

A *CrowdPulse* – *framework* de Musto et al., (2015) – quanto à identificação de discursos de ódio, apresentou uma abrangência nacional, uma vez que foram tomados dados de toda a Itália. Similar abrangência teve o Trabalho de Adamu et al., (2021), cujos autores partiram de uma Plataforma conhecida (CRISP-DM) e propuseram uma análise de emoções, tendo por base conteúdo textual, advindo do *Twitter*, em uma variação informal do inglês (*Pidgin*), construindo, para tanto, um *emotion dataset* nesta língua. No trabalho de Adamu et al., (2021) os autores partiram a campo com uma temática pré-definida, que seria a atuação do governo Nigeriano na distribuição de paliativos e kits de Socorro no combate à Pandemia da COVID-19.

Indo ainda mais ao encontro dos trabalhos com proposição da análise de sentimentos, sem, necessariamente, enfoque em enventos, apresenta-se o trabalho de Abdul-Rahman et al., (2021), onde é proposta uma *framework* para mineração e análise de dados, geolocalizados, do Twitter, utilizando de técnicas de Processamento de Linguagem Natural e Modelagem de Tópicos (LDA), no teste do modelo, em um estudo de caso feito no distrito de *Hung Hom*, com coleta que compreendeu um período de dez anos. A modelagem de tópicos, no trabalho em questão teve uma abordagem exploratória, não definindo, *a priori*, temas/assuntos e suas respectivas palavras-chave, mas deixando que os tópicos surgissem, justamente com a aplicação da modelagem de tópicos.

No trabalho de Abdul-Rahman et al., (2021) se utilizou a linguagem de programação *Python* para operacionalização de sua proposta, o que também fora feito no trabalho de Li et al., (2020) cujo objetio foi extrair opiniões de rede social sobre a

modificação genética em alimentos, com posterior aplicação da análise de sentimentos, tendo por fonte todos os *posts*/comentários sobre o assunto na conta da *People's Daily*, com a diferença de que não se utilizou o *Twitter*, mas sim o *Sina Weibo*, (*Twitter* chinês). Ainda que não assim definindo, se viu uma abordagem de Usuário Central, já no trabalho de Li et al., (2020).

Outro trabalho ainda mais relevante, quanto à abordagem de Usuários Centrais, foi o de Alkhatib et al., (2020), onde fora proposta uma *framework* para monitoramento da opinião pública em tempo real com base nos *tweets* de alto impacto de líderes públicos de opinião e seus seguidores, de modo a fornecer entendimento em profundidade e gerar relatórios do sentimento público. Os Usuários Centrias, no caso do trabalho de Alkhatib et al., (2020), foram justamente 52 líderes econômicos de países, a exemplo de seus presidentes, sendo que o teste do modelo se deu referente ao tema da guerra comercial entre China e Estados Unidos. A abordagem em questão fora reputada enquanto eficiente e escalável e foi, em certa medida, inspiração para o modelo que será apresentado posteriormente, no decorrer deste trabalho.

Passa-se, agora, à sessão onde serão descritos os procedimentos metodológicos, com maior detalhamento especificidade.

3 METODOLOGIA

A pesquisa, enquanto um processo sistemático e racional de resposta à problemas para os quais ainda não se têm respostas claras, compreende a cuidadosa seleção e utilização de métodos, técnicas e procedimentos (GIL, 2002), sendo, então, a metodologia, definida enquanto o “caminho do pensamento e a prática exercida na abordagem da realidade” (MINAYO, 2002, p.16).

Assim, os referidos métodos, técnicas e procedimentos, aqui delineados, tem o intuito de proporcionar a consecução dos objetivos elencados no capítulo 1 desta pesquisa. Esta seção da metodologia está seccionada em três partes, quais sejam: Enquadramento Metodológico, Objeto de Estudo e Modelo Proposto.

3.1 Enquadramento Metodológico

O tipo de pesquisa que aqui se propõe, está situado dentro do grande grupo das pesquisas em Ciências Sociais. Estas, que têm como objetivo imediato e último, respectivamente, a contribuição para a “aquisição do conhecimento e desenvolvimento do ser humano” e são “dirigidas para resolver problemas práticos” (RICHARDSON, 1999, p.16).

Desde logo, é apresentado, na figura abaixo, o enquadramento metodológico, de modo a facilitar, ao leitor, a compreensão do tipo de pesquisa que aqui se propõe realizar, conforme os métodos escolhidos para tanto, e, após a Figura 5, são trazidas as explicações pertinentes.

Figura 5 - Enquadramento Metodológico

Fonte: dados de pesquisa (2022)

Quanto à Natureza, trata-se de pesquisa aplicada. Na pesquisa aplicada, as razões principais de se pesquisar estão fundadas na melhoria prática de algo, tornando-o mais eficiente ou eficaz, em detrimento às razões puramente de ordem intelectual, que denotam o que se chama de pesquisa básica (GIL, 2002).

É defendido, no entanto, que tais conceitos de definição da pesquisa, quanto à natureza, na prática, não sejam mutuamente exclusivos, haja vista não haver óbice à pesquisas práticas encontrarem novos princípios científicos ou pesquisas básicas possuírem aplicação prática, de imediato (GIL, 2002).

Quanto aos objetivos, entende-se esta pesquisa como sendo de cunho predominantemente exploratório, tendo em vista o propósito de ir “a campo” buscar o(s) tópico(s) mais predominante(s) (e seu(s) sentimento(s) associado(s)), em discussão pelos *netizens* da cidade *lócus*, intersectando, ainda, aspectos de um estudo de caso – quanto aos procedimentos – conforme observado como sendo algo usual, em pesquisas exploratórias (GIL, 2002), o que ratifica o enquadramento aqui comentado.

Há intersecção, contudo, com conceitos de pesquisa descritiva, também, tendo em vista o intento de retratar as relações entre variáveis (GIL, 2002), quais sejam, a nível macro, o(s) tópico(s) identificado(s), *a posteriori*, a partir do conjunto de dados, e a polaridade do sentimento associada, enquanto negativo, positivo ou neutro.

Gil (2002, p.42) arrazoa que nas pesquisas descritivas se “têm por objetivo estudar as características de um grupo”, as quais, neste caso, é o próprio grupo de *netizens,* tendo por fonte de dados suas opiniões exaradas na rede social *Twitter* e seus sentimentos referentes ao(s) principal(is) tópico(s) identificado(s).

As pesquisas exploratórias e descritivas são as que “habitualmente realizam os pesquisadores sociais preocupados com a atuação prática” (GIL, 2002, p.42), afirmação que coaduna com a justificativa prática para realização do presente trabalho.

Quanto à abordagem, o trabalho em questão tendo em vista as análises propostas, que serão explicitadas adiante, é predominantemente quantitativo a julgar pelos métodos de *machine learning* empregados, os quais, fundamentalmente, são traduzidos em métodos estatísticos, tanto para análise do sentimento, quando para a identificação dos tópicos, por meio da LDA. Contudo, há intersecções com abordagens qualitativas, a exemplo, justamente, da análise e rotulação dos tópicos identificados no processo de LDA.

Quanto aos Procedimentos, trata-se de um teste de aplicação de modelo, que intersecta a metodologia de estudo de caso, a ser feito tendo por lócus a cidade de Maringá. O estudo de caso tem por objetivo o conhecimento amplo e detalhado de um ou poucos objetos, cujos resultados são apresentados de maneira aberta, usual a

trabalhos de cunho exploratório (GIL, 2002). Ainda, também pode ser entendido enquanto um “trabalho de caráter empírico que investiga um dado fenômeno dentro de um contexto real contemporâneo por meio de análise aprofundada de um ou mais objetos [...] possibilita amplo e detalhado conhecimento sobre o fenômeno” (CAUCHIK-MIGUEL et al., 2012, p.131).

Contudo, quanto aos procedimentos, o encaixe mais adequado da metodologia é condizente com a proposição de *Framework* conceitual, sendo, neste caso, para extração e análise de dados do *Twitter*. O método de construção de *Framework* advém da técnica de abordagem por processos, de Platts e Gregory (1990), com consolidação por Lima e Costa (2004), no âmbito da Engenharia de Produção, que compreende

> o desenvolvimento de uma abordagem prescritiva, que operacionaliza um conjunto de conceitos, através de um processo estruturado, e com instrumentos de coleta de dados, dinâmica e critérios de avaliação (CAUCHIK-MIGUEL et al., 2012, p.200)

Parte-se da literatura revisada, e das oportunidades de pesquisa vislumbradas, por meio da qual é feito o modelamento de processo que responda, operacionalmente, às necessidades de acesso aos dados que aqui se vislumbram. Tal qual arrazoado por Lima e Costa (2004, p.34), referente à construção de *Framework*, trata-se da “articulação de uma teoria que orienta a prática” tendo em vista os “relacionamentos das atividades que compõe o processo”.

Na esteira da construção de *Framework/*Modelo – apresentado adiante – é necessário a observação referente ao Instrumento de Pesquisa. Entende-se por Instrumento de Pesquisa, aquilo que “é utilizado para a coleta de dados” (RUDIO, 2007, p.114). Assim, desde já se salienta que o modelo proposto, é o próprio Instrumento de pesquisa do presente trabalho, na medida em que, por meio dele é que são coletados os dados para posterior análise.

Ainda, no ínterim da coleta de dados, arrazoa-se, uma vez mais, quanto às oportunidades da utilização dos dados de redes sociais, (neste caso, *tweets*) os quais, enquanto fonte secundária[13], apresentam vantagens sobre outras formas de coleta de dados, tais como as *surveys* tradicionais ou entrevistas, na medida em que sua coleta

[13] Acesso em 16 de Abril de 2022. Disponível em: https://saylordotorg.github.io/text_emarketing-the-essential-guide-to-online-marketing/s21-03-primary-and-secondary-research.html.

são menos custosos, mais céleres, não provocados pelo pesquisador, sendo, assim, menos sujeitos à enviesamentos (ALIZADEH; SARLAR; BIRGPUME, 2019, EL-DIRABY; SHALABY; HOSSEINI, 2019, LI et al., 2020). Tendo sido feito o enquadramento metodológico, passa-se, na sequência, à construção do modelo.

3.2 Construção do Modelo

Nesta seção, se intenta atingir o objetivo de proposição de um modelo para extração de dados, tendo por fonte a rede social Twitter, bem como a análise destes dados, sendo que, a construção do referido modelo e os argumentos para cada uma de suas etapas são aqui explicados, junto da metodologia. A apresentação do modelo terminado será feita no capítulo de Resultados.

Busca-se, também, e por conseguinte, o atendimento aos objetivos específicos desta dissertação, por meio da apresentação dos passos utilizados para elaboração do código de extração (a), identificação do(s) tópico(s) proeminente(s) (b), extração de dados temática (c) e aplicação dos modelos de Análise de Sentimento (d).

Começa-se tratando da macroestrutura da *Framework* proposta. Se buscou a utilização de modelos de *Frameworks* já consolidados, as quais ajudarão na disposição das etapas e passos metodológicos, quais sejam a *CUP Framework* (FAN; GORDON, 2014) combinada com a *ECCO Framework* (KAZMAIER; VUUREN, 2020).

Utilização da *CUP Framework* tem por objetivo, justamente, dados de Redes Sociais – o que concatena com o que aqui se intenta – possuindo três etapas principais que são: *Capture* (Captura de dados relevantes, de redes sociais, extraindo o que for pertinente), *Understand* (Entendimento dos dados capturados, removendo dados sem qualidade ou ruidosos) e *Present* (Apresentação das informações, posteriormente, de modo inteligível) (FAN; GORDON, 2014). Estas etapas serão tratadas neste trabalho, majoritariamente, em português por: Captura, Entendimento e Apresentação, respectivamente.

A opção pela utilização da *CUP Framework* está amparada, por exemplo, por Kankanamge et al, (2020) os quais – em seu trabalho que propôs determinar a severidade de desastres por meio da análise de dados de redes sociais – argumentam que a utilização de uma metodologia mais abreviada (como é o caso da *CUP Framework*) ajuda na sintetização de metodologias extensas. Deste modo, com o

intuito de facilitar a visualização das etapas, optou-se pela adoção *CUP* enquanto modelo macro, para a metodologia proposta, a qual será complementada, em seu interior, com diretrizes e orientações trazidas pela *ECCO Framework*.

A ECCO Framework apresenta uma estrutura genérica para avaliação de dados textuais opinativos, não estruturados, cuja importante característica, é, justamente, a adaptabilidade a diferentes situações, sendo um modelo detalhado o suficiente para ser prontamente aplicado, sem perder seu cunho genérico de poder ser aplicável em contextos distintos, com as adaptações devidas, podendo, sua utilização, ser feita no todo ou em parte (KAZMAIER; VUUREN, 2020).

Uma vez mais, é explicitada a utilização de usuários centrais para coleta dos dados, considerando a disponibilidade de dados, por meio da utilização dos *tweets* direcionados a estes, representados pelas contas oficiais da prefeitura e do prefeito da cidade *locus*, na rede social *Twitter*.

A seleção destas contas, para operacionalização da coleta de dados, se dá em função de serem contas representativas da gestão da cidade. A nível macro, a abordagem de utilizar Usuários Centrais foi feita por Alkhatib et al., (2020), onde os autores, com base numa fórmula considerando o somatório de competência e popularidade, identificaram os líderes de opinião em política, economia e educação de modo a extrair destes e de seus seguidores, sentimentos acerca da guerra comercial entre EUA e China.

No caso da presente dissertação, tendo em vista se tratar de *lócus* micro – em detrimento ao nível mundial, do trabalho de Alkhatib et al., (2020) – a identificação dos líderes de opinião é feita tendo em vista, sempre a disponibilidade de contas oficiais, no *Twitter,* do Prefeito (1) e da Prefeitura (2) da cidade objeto de estudo, sendo preferível, para coleta de dados, que a haja as duas contas disponíveis, entendendo que os *tweets* direcionados a estas duas contas já teriam o objetivo de apontar problemas, sugestões, situações, reclamações ou elogios a respeito do que concerne a vida dos cidadãos da cidade, usuários do *Twitter*, diretamente aos interessados e responsáveis, com poder de decisão. Ressalta-se, ainda, que a escolha destes usuários centrais não é um elemento que entra como etapa ou passo, do modelo proposto, por ser uma característica anterior à operacionalização do modelo, ou seja, é condição prévia a ser considerada antes de se realizar as demais etapas, junto da própria escolha da cidade *lócus* na qual o modelo será aplicado.

O Modelo subdividir-se-á em duas Macro-Etapas, as quais correspondem a dois ciclos completos de Captura (*Capture*), Entendimento (*Understand*) e Apresentação (*Present*) dos dados, de modo que seguir-se-á esta lógica para seccionamentos seguintes, nesta seção secundária de Construção do Modelo.

3.2.1 Captura_1

A primeira etapa do modelo começa com a Coleta ou Captura dos dados, os quais, são extraídos do *Twitter.* Desde já, justifica-se a escolha da fonte de dados *Twitter* por se tratar de Rede Social rica em dados, com um grande volume de publicações sendo produzidos constantemente, por uma grande quantidade de usuários (ALSAEDI; BURNAP; RANA, 2017, HASNAT; HASAN, 2018).

Ademais, a vantagem do *Twitter,* enquanto uma Rede Social do tipo *Microblog*, é a de que os usuários têm uma limitação de caracteres a cada postagem que, atualmente, é de 280 caracteres por *tweet* – no passado era limitado à 140 caracteres, parte em função da sua origem enquanto um serviço baseado em SMS[14] – os quais, de certo modo, “forçam” o usuário a ser sucinto na mensagem a ser expressa, indo direto ao ponto (GASCO et al., 2019).

Ainda, trata-se de Rede Social de cunho mais “aberto”, que coaduna com a ideia de *open data*, disponibilizando, ainda, uma *Application Programming Interface* (API) simples (KANKANAMGE et al., 2020) que permite a pesquisadores e demais interessados a requererem os dados (*tweets* e *metadata*) diretamente à *Back-End* da plataforma. O acesso, operacionalmente, é feito por meio do *scraping*[15] ou “raspagem” de dados, que pode ser entendido como um modo automático de “copiar e colar” as informações públicas, contidas na rede, feita com base em critérios ou parâmetros definidos no Código fonte do programa “raspador”.

Deve, também, se levar em conta, a título de comparação, que outras plataformas como *Facebook*, por exemplo – não obstante ser a rede social mais utilizada no Brasil[16] – já não são tão receptivas ao acesso de terceiros a seus dados,

---

[14] Acesso em 5 de Fevereiro de 2022. Disponível em: https://developer.twitter.com/en/docs/counting-characters.

[15] Acesso em 6 de Agosto de 2022. Disponível em: https://www.gocache.com.br/seguranca/o-que-e-web-scraping-para-iniciantes/

[16] Acesso em 6 de Agosto de 2022. Disponível em: https://gs.statcounter.com/social-media-stats/all/brazil

por meios automatizados, havendo, por padrão, a proibição de qualquer tipo de *scraping* na plataforma[17], tanto que, nenhum dos trabalhos revisados tem o *Facebook* como fonte de dados, seguindo uma lógica de extração automatizada, como aqui se propõe. O *Facebook*, inclusive, expressa explicitamente que a coleta automatizada de dados é proibida, exceto se houver expressa permissão escrita da plataforma para isso[18], sendo que a referida permissão não é intuitivamente acessível, e, ainda que acessível e obtida, a coleta automática é confinada, por padrão, para somente ser visualizada na internet[19]. Não bastasse isso, há página própria do *meta/Facebook* indicando o combate ao *scraping* na plataforma[20].

Assim, a escolha do *Twitter* como fonte de dados é corroborada, também, pela sua utilização frequente como *data source* da maior parte dos trabalhos revisados – a exemplo de trabalhos recentes como os de Alkhatib et al., (2020), Sánchez-Ávila et al., (2020), Abdul-Rahman et al., (2021), Fan, Jiang e Mostafavi, (2020), Wang et al., (2020), Sakurai e Adu-Gyamfi, (2020), Rahimi-Golkhandan, Garvin e Wang (2021), Yuan et al., (2021), Adamu et al., (2021) e Marzouki et al., (2021) – fundamentando os motivos de sua escolha.

O que, sobretudo, interessa, aos propósitos deste trabalho, são os conteúdos textuais advindos das postagens feitas pelos usuários do *Twitter*, as quais são tidas como *tweets*. Contudo, ainda que se estabeleça a primazia pelo conteúdo textual dos *tweets,* intenta-se a extração de todas as *features* e metadados dos *tweets* de modo a ter a possibilidade de realizar outras análises, que não apenas as análises LDA e *Sentiment Analysis,* ao encontro do que recomenda a literatura (WANG et al., 2015, ANDREWS et al., 2016, ALKHATIB; EL BARACHI; SHAALAN, 2019, GASCO et al., 2019, EL-DIRABY; SHALABY; HOSSEINI, 2019, ALIZADEH; SARLAR; BIRGPUME, 2019, ALKHATIB et al., 2020, ADAMU et al., 2021, MARZOUKI et al., 2021).

Ainda, com relação à redes sociais, como um todo, e presença das organizações públicas, fora verificado, no último relatório “Pesquisa TIC Governo Eletrônico 2021”[21], do Comitê Gestor da Internet no Brasil (CGI.br), ainda que não

[17] Acesso em 17 de Janeiro de 2022. Disponível em: https://www.facebook.com/robots.txt.
[18] Acesso em 6 de Agosto de 2022. Disponível em: https://www.facebook.com/robots.txt
[19] Acesso em 6 de Agosto de 2022. Disponível em: https://www.facebook.com/apps/site_scraping_tos_terms.php
[20] Acesso em 6 de Agosto de 2022. Disponível em: https://about.fb.com/news/2021/04/how-we-combat-scraping/.
[21] Acesso em 6 de Agosto de 2022. Disponível em: https://www.cgi.br/media/docs/publicacoes/2/20220725170710/tic_governo_eletronico_2021_livro_eletronico.pdf.

citando explicitamente o *Twitter*, um aumento na presença de prefeituras nas redes sociais de 82% em 2019 para 94% em 2021, e isto é algo que é importante para a operacionalização do modelo, tendo em vista a utilização da presença e conta da prefeitura no *Twitter*, enquanto um usuário central para a focalização da coleta de dados.

Tendo sido definida a fonte, é necessária a compreensão de como abstrair os dados produzidos da citada fonte. A extração de dados do *Twitter* é realizada por meio do acesso à *Application Programming Interface* (API), disponibilizada pela própria plataforma, para desenvolvedores que queiram acessar os dados de maneira programática[22]. Para tanto, contudo, é necessário que o usuário possua uma conta como usuário da Rede Social, e, então, solicite acesso como desenvolvedor, junto à plataforma, justificando seus objetivos de acesso diferenciado, de modo a conseguir as API *Keys* para fazer requisições ao *back-end* do *Twitter*[23].

Conseguido o acesso às API *Keys* – cumprindo o **passo 1** do Modelo – para a operacionalização da coleta, de maneira programática, é necessário a escolha de um editor de código, compatível, neste caso, com a linguagem de programação *Python* 3, escolhida para implementação do algoritmo de extração, sendo este o **passo 2**.

O *Python* é, atualmente, uma das linguagens de programação mais populares do mundo[24], e, parte significativa de sua popularidade acredita-se ser devida ao fato de se tratar de opção simples de aprender e barata (ABDUL-RAHMAN et al., 2021), verificando-se, comumente, na literatura, a implementação das estratégias de coleta/*scraping* de dados fazendo, justamente, uso desta linguagem (POORAZIZI; HUNTER; STEINIGER, 2015, ALKHATIB; EL BARACHI; SHAALAN, 2019, ABDUL-RAHMAN et al., 2021, LI et al., 2020, ADAMU et al., 2021, YUAN et al., 2021).

O acesso à *Twitter API* é o **passo 3** do Modelo e é feito por meio da Criação de um *script* programático, via editor de código escolhido, onde a comunicação com à API é feita, no início, por meio da identificação do usuário, atribuindo as chaves de acesso (API *Keys*) à variáveis de autenticação. Assim, se consegue o acesso ao *back-end* da plataforma.

---

[22] Acesso em 5 de Fevereiro de 2022. Disponível em: https://developer.twitter.com/en/docs/twitter-api.
[23] Acesso em 5 de Fevereiro de 2022. Disponível em: https://developer.twitter.com/en/docs/twitter-api/getting-started/getting-access-to-the-twitter-api.
[24] Acesso em 27 de Janeiro de 2022. Disponível em: https://www.tiobe.com/tiobe-index/.

Parte-se, então, ao **passo 4**, onde, as primeiras ações dizem respeito à importação de pacotes *Python* que facilitam a extração e tratamento dos dados acessados, sobretudo, quanto ao próprio acesso à API do *Twitter* e o "manuseio" dos *tweets* extraídos. A utilização destes Pacotes, parte, uma vez mais, do pressuposto colaborativo do próprio *Python*, enquanto uma linguagem *open-source*, na qual a comunidade desenvolvedora partilha seus códigos, tornando mais eficiente a construção de projetos, de modo que não se tenha que começar um *script* sempre "do zero".

Para a extração fática dos *tweets* e seus metadados – tratando, agora, do **passo 5** – segundo os propósitos do presente artigo, alguns parâmetros precisam ser discriminados. Primeiramente, desde logo, define-se que o modelo aqui proposto segue a lógica de uma *Framework* Não Temática, sem definição de assunto ou domínio prévio – tal qual trabalho de Abdul-Rahman et al., (2021) – ou seja, a extração de dados é feita, inicialmente, sem um tema definido a *priori,* deixando que este, ao revés, surja a *posteriori,* através da análise do *dataset* inicial, com este fito.

Isto posto, o desafio seguinte é estabelecer um mecanismo de extração de *tweets* adequado, hábil a compor um *dataset* suficientemente situado, geograficamente, na cidade *lócus* de aplicação do modelo.

Assim, uma abordagem inicial possível seria a busca por todos os *tweets* geolocalizados na região da cidade *lócus*, ou seja, *tweets* com as exatas coordenadas geográficas no momento da postagem, condizentes à cidade escolhida, num raio de 10 quilômetros, ou, na falta destes, *tweets* advindos de usuários habitantes da referida cidade, segundo constante em seu perfil.

Entretanto, verificou-se (em testes de coletas preliminares, feitos entre os dias 10 e 11/01/2022) que esta abordagem acabava trazendo dados muito ruidosos, visto que estariam sendo acessados, em grande medida, *tweets* sem qualquer relevância prática ao contexto da cidade, uma vez que eram abstraídas conversas de usuários sobre assuntos bastante específicos, como comentários sobre esportes, filmes, conversas entre usuários sobre tópicos de interesses individuais.

Assim, a estratégia reputada como adequada, de modo a obter um *dataset* menor, porém mais focado – e, assim, mais sustentável – é a de extrair, num primeiro momento, somente *tweets* direcionados (sejam respostas/*replys* ou não) às duas contas oficiais – seus "@" – dos usuários centrais da cidade na qual o modelo será aplicado, qual sejam, a conta da prefeitura, e a conta do prefeito, no *Twitter*.

É feita esta escolha por entender que os *tweets* direcionados a estas duas contas já teriam o fito de apontar problemas/sugestões/situações/reclamações/elogios a respeito do que concerne a vida dos *netizens* da cidade, diretamente aos interessados e responsáveis, com poder de decisão.

Ademais, deste modo, se está acenando ao que fora observado na revisão de literatura, de optar por abordagens mais sustentáveis/focadas, vislumbrando, ainda, a utilização de Usuários Centrais, numa estratégia razoável ao propósito deste trabalho.

Como parâmetros de extração, é feita, ainda, a opção por não extrair *retweets* (RTs), logo na raiz do *script*, visto que, na prática, se referem à opinião/comentário de outra pessoa sobre algum assunto (GASCO et al., 2019), não necessariamente refletindo o posicionamento do usuário que retuitou. Este filtro também foi justificado por Li et al., (2020), no qual, *forwaded messages* – similar ao *retweet* – na Rede Social *Weibo* (*Twitter* chinês) também foram desconsideradas por não refletirem as emoções próprias do usuário que as "retuitou". Ademais, ratifica-se que a filtragem de *retweets* encontra respaldo na literatura (ANDREWS et al., 2016, JOSEPH et al., 2017, EL-DIRABY; SHALABY; HOSSEINI, 2019, GASCO et al., 2019, ALKHATIB et al., 2020, SÁNCHEZ-ÁVILA et al., 2020, LI et al., 2020, LIU; TENG; GONG, 2021).

Ressalta-se que a abordagem geolocalizada de extração de *tweets*, por mais que não utilizada neste primeiro momento, é uma das estratégias empregadas para a aquisição do *dataset* para posterior Análise de Sentimentos, conforme será novamente comentada, adiante.

É, também, utilizado filtro de não absorver *tweets* advindos diretamente das contas do Prefeito ou da Prefeitura, caso estes, porventura, estivessem presentes.

Destes 5 passos trazidos neste tópico, o *output* será um *corpus* de documentos, planilhados, que servirão de *input* à próxima etapa.

3.2.2 Entendimento_1 e Apresentação_1

Desde logo, justifica-se o presente tópico tratar tanto da parte de Entendimento, como da parte de Apresentação*,* desta Macro-Etapa 1, pelo caráter iterativo e de reverificação das etapas que compreendem os passos de 8 a 12, sobretudo de 8 a 10.

O primeiro passo desta etapa é o **passo 6,** que diz respeito à importação do *corpus* comentado logo acima, para dentro do *software Orange Data Mining*[25], que se trata de uma solução *open source* para análise de dados, tendo por base de sua construção, justamente, a linguagem *Python*. Uma vantagem de se utilizar o *Orange Data Mining*, diz respeito às inúmeras funcionalidades interativas para análise e visualização dos dados, as quais podem ser implementadas sem necessidade de conhecimentos de programação, na medida em que se trata de uma ferramenta de programação visual, acessível, deste modo, a um público maior de usuários, razão que levou à sua escolha para execução das etapas de Entendimento e Apresentação, do modelo.

Dentro do *Orange*, passa-se ao **passo 7**, que diz respeito à Programação do Pré-Processamento. O pré-processamento é necessário, antes de performar as análises, na medida em que os dados extraídos vêm ruidosos, sendo indispensável sua limpeza.

Deste modo, é selecionada, do *corpus*, a variável do texto do *tweet*, e nesta, apoiando-se na literatura revisada, é que se determina a realização dos seguintes procedimentos:

- colocar todo o texto em letra minúscula (JOSEPH et al., 2017, ADAMU et al., 2021, MARZOUKI et al., 2021);
- remover *links* e URLs (ALIZADEH; SARLAR; BIRGPUME, 2019, ALKHATIB et al., 2020, SÁNCHEZ-ÁVILA et al., 2020);
- "tokenizar" os *tweets*, que equivale a quebrar um documento (*tweet*) a nível de palavra (COSTA et al., 2018, JAIN; KUMAR, 2018, ALKHATIB et al., 2020);
- normalizar com "steemização", que é trazer os tokens ao seu radical (ALSAEDI; BURNAP; RANA, 2017, JAIN; KUMAR, 2018, COSTA et al., 2018, EL-DIRABY; SHALABY; HOSSEINI, 2019; ALKHATIB; EL BARACHI; SHAALAN, 2019);
- remover *stop words,* que é o processo de remover palavras sem muito significado para análise, a exemplo de artigos e preposições (WANG et al., 2016, COSTA et al., 2018, JAIN; KUMAR, 2018, ALKHATIB; EL BARACHI;

[25] Acesso em 31 de Janeiro de 2022. Disponível em: https://dl.acm.org/doi/abs/10.5555/2567709.2567736.

SHAALAN, 2019, ALIZADEH; SARLAR; BIRGPUME, 2019, LIU; TENG; GONG, 2021, MARZOUKI et al., 2021);

- remover pontuação (JOSEPH et al., 2017, ALKHATIB et al., 2020, ADAMU et al., 2021).

Algumas considerações sobre a etapa de pré-processamento. Há uma iteração na etapa de remoção de *stop words*. Deve ser feita a visualização das palavras mais frequentes, após os filtros do pré-processamento, e, constando palavras que não agreguem significado analítico – ou mesmo pontuações não detectadas, pelo padrão *regexp* (expressões regulares), na etapa de remoção de pontuação – estas devem ser discriminadas em documento de texto que é alimentado ao *Orange*, indicando que seja removida pelo *widget* de pré-processamento, no *software.*

Ainda, orienta-se a definição de um intervalo de *N-grams* de 1 a 2. Tal intervalo tem o intuito de não ignorar possíveis expressões compostas (de até duas palavras), porventura frequentes no *corpus* de documentos.

O *output*, após o pré-processamento, é o *corpus* textual de documentos, limpo do ruído inicial, podendo, então, ser utilizado como *input* das análises que se planeja fazer.

A análise que se intenta, neste primeiro ciclo CUP, do modelo proposto, é a aplicação da *Latent Dirchlet Alocation* (LDA) para identificação do(s) tópico(s) latente(s) dado o *corpus* de documentos, encontrando respaldo na literatura, na medida em que é utilizada, com este mesmo fito, em outros trabalhos (WANG et al., 2016, JAIN; KUMAR, 2018, ABDUL-RAHMAN et al., 2021).

A LDA é um método estatístico/probabilístico – tratado como um método de *machine learning* não supervisionado, no campo de *Data Science* e *Artificial Intelligence* – que surgiu, no campo do aprendizado máquina, enquanto um meio – baseado nos modelos de *bag of words* – de clarificar relações estruturais/estatísticas latentes não só em um *Corpus* de documentos, mas também intra-documental (BLEI; NG; JORDAN, 2003).

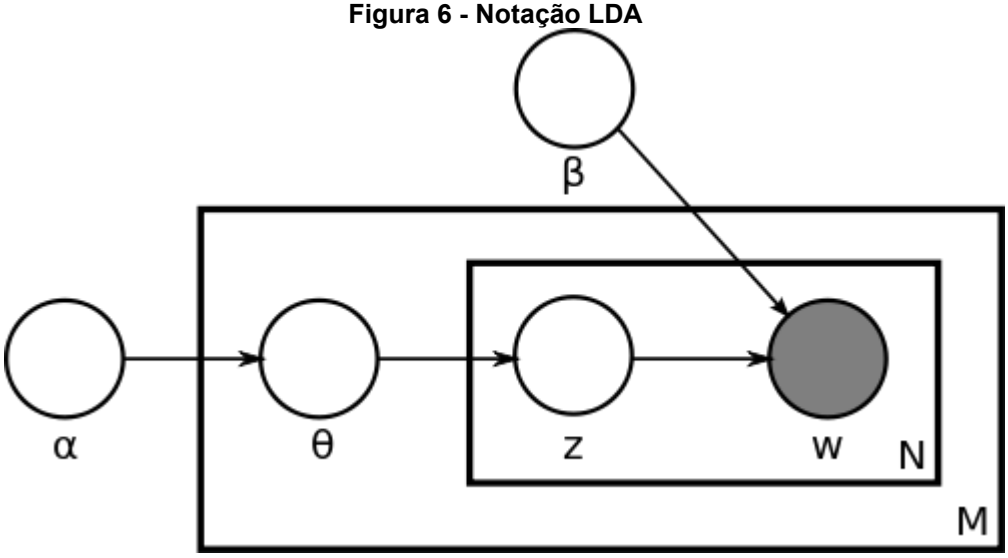
Na LDA cada documento que integra o *Corpus* é assumido enquanto representação de um ou mais tópicos presentes no conjunto de documentos, assmindo-sel, também, que cada tópico é formado por um mesmo conjunto frequente de palavras. A LDA utiliza um modelo gerador que simula a criação de documentos

com base nas relações observadas no *Corpus*, assumindo, assim, que as palavras sejam, então, geradas pelos tópicos (BLEI; NG; JORDAN, 2003). Acredita-se que a LDA, enquanto modelo generativo, tenha como uma de suas principais vantagens, a extensividade e modularidade, permitindo a inferência das relações existentes num *Corpus,* e aplicação em diferentes domínios (BLEI; NG; JORDAN, 2003).

O **passo 8** corresponde à etapa prévia à aplicação da *Latent Dirchlet Alocation* (LDA), que é a definição do número K de tópicos a serem considerados na operacionalização da análise. Por padrão, delimita-se K=10, e é então rodada a LDA, que é o **passo 9**.

A notação da LDA é explicada na sequência, por meio da Figura 8:

Figura 6 - Notação LDA



Fonte: Blei, NG e Jordan (2003)

M = indica o número de documentos no Corpus;

N = indica o número de palavras em um document;

α = é um parâmetro a *priori*, da LDA (*Dirichlet Prior*); é o parâmetro da probabilidade – a priori – nas distribuições de tópicos por documento. Quanto maior o seu valor, maior a probabilidade de cada documento contemplar uma mistura da maioria dos tópicos. Quanto menor seu valor, maior a probabilidade de cada documento contemplar apenas um ou dois tópicos específicos.

β = é o outro parâmetro a *priori*, da LDA (*Dirichlet Prior*), é o parâmetro da probabilidade – a priori – nas distribuições de palavras por tópico. Quanto maior seu valor, maior a probabilidade de cada tópico contemplar uma mistura

da maioria das palavras. Quanto menor seu valor, maior a probabilidade de cada tópico contemplar apenas algumas palavras;

θ(*i*) = é a distribuição de tópicos (*j*) por documento (*i*);

Z(*ij*) = é a atribuição de tópico para cada palavra W(ij), ou o tópico da j- ésima palavra no i-ésimo documento; em outras palavras, é usado para anotar cada tópico atribuído a cada palavra, fazendo com que cada documento seja uma mistura desses tópicos;

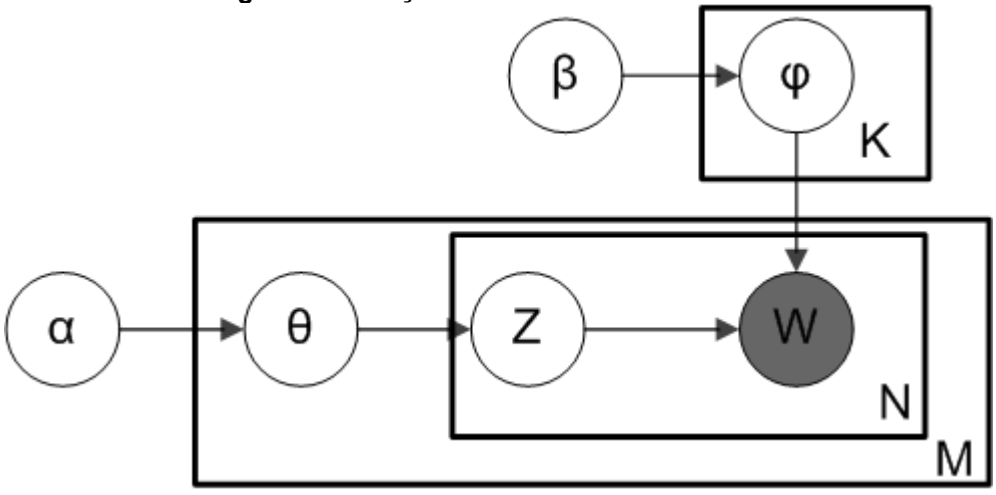
W(*ij*) = é a j-ésima palavra no i-ésimo documento.

Para o propósito de identificação de tópicos latentes, num dado *Corpus* de documentos, a LDA realiza um processo "reverso", que é, justamente, a suposição de como novos documentos seriam gerados. Tal processo é denominado *generative process*.

Neste dado *Corpus* de documentos, se propõe descobrir a distribuição dos K tópicos, para cada documento do *dataset*, indicando, também, a distribuição das palavras para cada um dos tópicos. O LDA faz isso partindo do princípio de que, como já dito, cada tópico é formado por um mesmo conjunto frequente de palavras.

A notação do modelo resultante é dada abaixo, pela Figura 9, contemplando o elemento K, que se refere ao número de tópicos.

Figura 7 - Notação LDA resultante/*smoothed*



Fonte: Blei, NG e Jordan (2003)

O elemento φ de K, pode ser entendido em complemento ao parâmetro θ, do modelo resultante. Ambos denotam matrizes, sendo θ formada, por *i* documentos, cada qual sendo associada distribuição dos tópicos *j,* enquanto φ exprime as palavras W*j* para cada tópico *i*.

Da visualização de θ e φ, juntamente ao controle dos parâmetros dados por α e β, é que se verifica a razoável definição – ou não – da quantidade K de tópicos, partindo da premissa de que a definição de um K adequado denota documentos contemplando apenas um ou poucos tópicos específicos, e tópicos formados por um conjunto pequeno de palavras representativas.

Na prática, a verificação da adequação do valor definido para K, é feita no **passo 10**, é dentro do próprio *Orange,* onde se observa, de maneira gráfica, a similaridade dentro do conjunto de dados, considerando a aplicação da LDA.

Os **passos de 6 a 9** estão compreendidos dentro da raia Entendimento, deste primeiro ciclo CUP, enquanto a partir do passo 10, passa-se à raia da Apresentação.

Tendo definido um padrão para rodagem inical de K=10, o que é feito, na sequência, com o **passo 10**, é um processo iterativo de visualização da distribuição dos 10 *clusters*, num plano bidimensional, através da *Multidimensional Scaling* (MDS) (WICKELMAIER, 2003). Por meio da MDS é possível visualizar, graficamente, a distribuição e distância entre os tópicos, representados por pontos, para os quais fora aplicado a LDA (ou outro algoritmo de modelagem de tópicos).

A visualização, via MDS permite, ainda, por meio de um esquema de cores, identificar a relevância ou "presença" de cada tópico dentro do conjunto de dados avaliado. Quanto mais próxima a distância entre dois pontos/tópicos, maior a similaridade entre eles, na medida em que é possível que se tratem, na prática, do mesmo tópico. Quando relações como esta são observadas – de tópicos avizinhados – deve ser reduzido o valor de K de modo a ser condizente com a distribuição observada, intentando um número ótimo de tópicos, onde estes sejam distintos entre si.

Os outros dois passos, referentes à parte de Apresentação, deste primeiro ciclo CUP, são os **passos 11 e 12** os quais traduzem um processo de visualização das *keywords*, ou termos mais frequentes para cada tópico processado via LDA, o qual, conjugado com a visualização dos documentos mais representativos de cada tópico, permitem verificar, desde logo, os assuntos preponderantes tratados em cada tópico.

Na LDA é dado um *score* de pertencimento à cada documento (*tweet,* neste caso), frente a cada tópico, ou seja, diferente de outros métodos de clusterização, a presença de um documento em um tópico não determina sua exclusão em outro(s) tópico(s).

A visualização, no *console* do *Orange*, dos **passos de 10 a 12** podem corresponder às etapas de *feedback* ao usuário, conforme preconizada por Kazmaier e Vuuren (2020), onde se comenta, inclusive, que os *feedbacks* sejam fornecidos ao longo do processo, e de maneira iterativa.

Caso as etapas compreendidas pelos **passos de 8 a 12** tenham retornado resultados significativos, com tópicos razoavelmente delimitados, retorna-se à raia do Entendimento, na qual, no **passo 13**, os tópicos são rotulados, segundo sua temática principal. É importante que este passo seja feito de maneira a considerar não somente as principais palavras de cada tópico, mas também os principais documentos (*tweets*) de cada tópico, permitindo, assim, uma análise mais embasada.

Na sequência, tem-se o **passo 14**, onde são abstraídas as principais *keywords* que representam os tópicos, as quais serão usadas de *input* para o processo de Captura da Macro-Etapa 2. Este processo tem o fito de facilitar a coleta de dados de maneira tematicamente focalizada, tendo em vista a utilização dos termos/*tokens* do próprio *corpus* de documentos enquanto indicador das palavras mais significativas e com potencial de retornar o maior número de *tweets* representativos.

3.2.3 Captura_2

É feita a explicação, desde já, de que os requisitos de acesso à *Twitter API* e ações discriminadas nos **passos de 1 a 5** continuam os válidos, para esta segunda etapa de Captura, mas que, por razões de sintetização da programação visual do modelo, não são repetidos.

Para esta nova captura dos dados, reforça-se o quanto já dito de que utilizar-se-á a ferramenta de busca, na *Twitter API*, de maneira tematicamente focalizada. As principais palavras do(s) tópico(s) mais proeminente do *corpus* analisado, na Macro-Etapa 1, devem ser combinadas de duas formas diferentes, através de dois *scripts* construídos para extração, cada qual seguindo uma lógica distinta. Os referidos *scripts* são representados pelos **passos 15 e 16**, e seguem a seguinte sistemática:

Lógica de extração do **passo 15**: Combinando, com operadores booleanos, dois eixos de menção (AND entre eixos e OR intraeixo):

i. Conta no *Twitter* do Prefeito da cidade lócus OR Conta no *Twitter* da Prefeitura da cidade lócus;

ii. Principais palavras componentes do(s) tópico(s) mais proeminente(s).

Lógica de extração do **passo 16**: Combinando, com operadores booleanos, dois eixos de menção (AND entre eixos e OR intraeixo):

i. Principais palavras componentes do(s) tópico(s) mais proeminente(s).

ii. Palavras que remetam ao contexto das cidades, tais como: “prefeitura”, “prefeito”, “cidade”, “nome da cidade”

O **passo 15** segue, grosso modo, a mesma lógica de extração explicada na seção Captura_1, devendo ser, agora, contudo, focada tematicamente, por meio da intersecção das *keywords* principais, abstraídas da identificação do(s) tópico(s) proeminente(s), via LDA.

Na Lógica de extração do **passo 16** deve ser utilizado, ainda, de modo a restringir os resultados à cidade lócus do estudo, o filtro *geocode*, por meio do qual, inserindo as coordenadas geográficas da cidade de aplicação do modelo, e estipulando um *range* de até 10km de raio, será possível extrair somente *tweets* geolocalizados da cidade escolhida, seja pela localização ativada no momento da postagem, ou, no caso da indisponibilidade desta, pela identificação no perfil do usuário. O *range* de 10km de raio é aproximado, e que, a depender de onde se está aplicando o modelo, deve ser ajustado para mais, no caso de cidades geograficamente mais extensas em área, ou diminuído, caso contrário.

A ideia, no **passo 16,** é a de buscar *tweets* que sejam relevantes sobre o assunto do(s) tópico(s) principal(is) identificado(s), mas que não façam menção direta nem ao Prefeito e nem à Prefeitura, mas que expressassem a voz dos *netizens* sobre o(s) tópico(s), ainda que, em tese, restritos apenas a sua lista de contatos/seguidores.

Em ambos os modos deve ser mantida a utilização da filtragem de RTs, nativa. Em ambos os modos deve ser mantido o filtro de não absorver *tweets* vindo das contas do Prefeito ou da Prefeitura.

Importante destacar, também, as limitações da versão gratuita da *Twitter API,* quais sejam o limite cronológico de acesso pretérito a *tweets*, que varia num intervalo de 6 a 9 dias anteriores a data da *query*[26]. Este fato leva à necessidade de rodar os *scripts* datas distintas, de modo a cobrir o espaço de coleta desejado.

[26] Acesso em 5 de Fevereiro de 2022. Disponível em: https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/guides/standard-operators.

Destes *scripts* de iteração na *Twitter API*, para extração de dados, o *output* é um *dataset* extraído conforme os parâmetros comentados. Referido *Dataset* servirá de *input* à próxima etapa.

3.2.4 Entendimento_2

Inicia-se a seção Entendimento da Macro-Etapa 2, referente à construção do modelo, similarmente ao que foi feito na mesma parte da Macro-Etapa 1, ou seja, o que se deve fazer, neste momento, é a importação do *Dataset* temático, obtido através dos procedimentos anteriormente mencionados, para dentro do *software Orange Data Mining*, realizando, assim, o **passo 17**.

Tal qual o passo 17 remonta ao passo 6, da Macro-Etapa 1, o **passo 18**, da programação do Pré-Processamento remonta ao passo 7, homônimo, na medida em que os pressupostos de necessidade da diminuição do ruido, limpeza dos dados dentre outros cuidados necessários de serem tidos, antes da análise do *Dataset*, continuam válidos. Os mesmos procedimentos de limpeza, filtragem e normalização, descritos no passo 7, devem ser aplicados, aqui no passo 18, de modo que não serão novamente elencados. Deve ser, contudo, arrazoado, que, conforme recomendação de Kazmaier e Vuuren (2020), buscar-se-á a otimização dos resultados de classificação dos modelos por meio da iteração nesta etapa de pré-processamento, alterando as configurações de *cleansing* de dados, tendo em vista o alcance dos melhores parâmetros de filtragem para a obtenção dos melhores resultados dos modelos.

A exceção à aplicação do pré-processamento, tal qual o passo 7, diz respeito à não remoção de toda a pontuação dos documentos, uma vez que pontuações como interrogações (?) e exclamações (!) podem ser úteis na diferenciação da polaridade de sentimento, sobretudo em estratégias não supervisionadas como o VADER (OAD et al., 2021).

Havendo obtido, como resultado, um *Dataset* limpo, *post-processing*, este servirá de *input* às estratégias de análise que aqui são propostas, representadas pelas ações compreendidas nos passos de 19 a 23.

Duas das principais análises aqui propostas, são trazidos nos **passos 19 e 20**, ambas se referindo à análise de sentimento, cada uma, contudo, seguindo lógicas distintas. Enquanto a análise trazida pelo passo 19 é supervisionada, ou seja, implica

no treinamento de modelo(s) de *machine learning*, por meio da divisão do *dataset* num conjunto de treinamento/teste (geralmente segue-se a lógica de 80/20, respectivamente, como trazido por Kazmaier e Vuuren (2020), a lógica de análise do **passo 20** é não supervisionada o que implica na não existência do treinamento prévio do modelo, o qual tratará de encontrar padrões no conjunto de dados de *input*, de maneira exploratória. Diferenciações que podem ser feitas entre métodos supervisionados e não supervisionados, dizem respeito ao fato de, no primeiro, haver a necessidade de uma classificação/*labeling* prévia dos dados, enquanto no segundo, não há esse elemento, sendo encontradas relações no conjunto de dados, por meio de associação e clusterização (USHARANI, 2018).

Ainda, a lógica supervisionada implica no aprendizado estatístico, de modo a estimar valores ou características em um conjunto de dados, para prever a resposta do modelo utilizado, diante de uma nova observação (JAMES et al., 2013).

De um modo geral, a Análise de Sentimento pode ser definida enquanto a categorização das opiniões de um grupo diante de um dado assunto, produto, utilizando-se de uma fonte de dados textual para tanto, sendo as redes sociais, fontes excelentes considerando este intento (USHARANI, 2018). De todo modo, a Análise de Sentimento permite uma vasta gama de aplicações poderosas (KAZMAIER; VUUREN, 2020).

Mais especificamente a respeito da Análise de Sentimento seguindo a lógica supervisionada, do **passo 19**, tem-se, conforme já comentado, a necessidade de fazer um *labelling*, ou treinamento dos dados os quais servirão de *input* aos modelos, para classificação. Tal procedimento é feito como subdivisão do passo 19, no **passo 19.1**, onde é feita a divisão do *Dataset* em Treinamento e Teste, seguindo a proporção de 80/20, respectivamente, a exemplo do que fizeram Kazmaier e Vuuren (2020).

Necessário destacar que cada *tweet*, do referido conjunto de dados, é então classificado, de acordo com os conhecimentos do autor, tendo em vista uma das três possíveis categorias de sentimento, quais sejam positivo, negativo e neutro. Somente, então, sendo feita a divisão do *Dataset* conforme explicado no **passo 19.1**, é que é dada sequência no processo.

O *subset* de treinamento é então alimentado à quatro modelos para treinamento, quais sejam: Regressão Logística*, Naive Bayes, Support Vector Machines* e *Random Forest*, representados pelos **passos** que vão de **19.1.1 à 19.1.4**.

Passo 19.1.1 – Regressão Logística*:* trata-se de modelo estatístico capaz de, dado um *dataset* prévio de treinamento, ou preditores, modelar a probabilidade de uma variável Y pertencer a uma categoria em particular, podendo, as variáveis independentes, serem tanto contínuas como categóricas (sendo este último, o caso, aqui) (JAMES et al., 2013);

Passo 19.1.2 – *Naïve Bayes*: é um modelo para classificação que parte do princípio da independência entre os valores de K – não levando em conta seu ordenamento – para determinação de Y (JAMES et al., 2013), podendo ser adequado para classificação de variáveis discretas, como classificação de textos, com contagem de palavras[27];

Passo 19.1.3 – *Support Vector Machines* (SVM)*:* o modelo de classificação SVM trata de aumentar o *feature space*, ou o espaço das *n* dimensões onde constam as variáveis, de modo a poder lidar melhor com distribuições não lineares. Na SVM a ideia é encontrar a melhor *boundary,* ou secção, que divida os dados, corretamente, em classes/hiperplanos distintos (JAMES et al., 2013). O SVM tem sido considerado um bom classificador em uma distinta variedade de cenários (JAMES et al., 2013), dentre as quais, a detecção de *spam* em *e-mails*, a detecção de idiomas em textos e analises de sentimento[28], como é o caso presente.

Passo 19.1.4 – *Random Forest*: é definido por James et al., (2013) enquanto método de classificação tido dentro do conjunto dos métodos baseado em árvores, produzindo uma multiplicidade de árvores de predição, que são combinadas e, ao final, retornam uma predição consensual – para regressões – ou da predição da maioria dos ramos, o resultado do modelo – para classificação. Como vantagem de outros métodos de árvore, no *Random Forest* o *bootstrapping* trata de alimentar cada árvore com conjuntos de treinamentos diferentes, o que faz com que as árvores fiquem "decorrelacionadas" – *decorrelated* – reduzindo a variância e enviesamento no conjunto de dados (JAMES et al., 2013).

Os quatro modelos apresentados acima, do Passo 19.1.1 ao Passo 19.1.4, dado *input* de dados para treinamento dos modelos, são tidos como estratégias supervisionadas, a exemplo do que fizeram Adamu et al., (2021), onde foram

---

[27] Acesso em 2 de Maio de 2022. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

[28] Acesso em 2 de Maio de 2022. Disponível em: https://www.educative.io/edpresso/how-to-use-svms-for-text-classification

utilizados todos os modelos, de maneira supervisionada, que aqui se propõe – com exceção do *Naive Bayes* – para classificação de emoções apresentados num conjunto de *tweets* com lócus na Nigéria, tendo, em seu trabalho, o *Support Vector Machines* como melhor classificador (ADAMU et al., 2021).

Ainda, a escolha dos modelos em questão é feita muito em função da sua utilização em trabalhos revisados, do Portfólio Bibliográfico, conforme trazido nos parágrafos seguintes.

Análises próprias do sentimento, por meio supervisionado, a exemplo do que fizeram Li et al., (2020) utilizaram a Regressão Logística como modelo principal por entender que o modelo é maduro, com alta acurácia, apto a lidar com variáveis categóricas. Contudo, ambos os modelos – Regressão Logística e *Support Vector Machines* foram superados pelo modelo *Naive Bayes* na classificação de *tweets* como se tratando ou não de eventos de larga escala, a exemplo do que fizeram Alsaedi, Burnap e Rana (2017).

Já em seu trabalho que teve por objetivo utilizar dados de redes sociais para prever problemas no tráfego e gerar alertas para incidentes, Lu et al., (2018) obtiveram, dentre cinco modelos testados, o melhor resultado com uma adaptação do *Random Forest*, chamada *deep Random Forest,* obtendo um melhor (menor) MAPE e RSME que os demais métodos.

Jain e Kumar (2018) em seu trabalho cujo modelo tratava de tentar prever a ocorrência de doenças transmitidas por mosquito, na índia, fora aplicado classificadores para três classes, que seriam a de medo, sintomas e prevenção, sendo que, nos resultados, o SVM foi o algoritmo que mostrou melhor desempenho.

No já comentado trabalho de Alkhatib et al., (2020) os autores treinaram o SVM para que pudesse fazer a classificação de emoções, da análise de emoções, diferenciando entre cinco categorias de emoção: raiva, depressão, animação, felicidade e preocupação, sendo que o classificador obteve um desempenho de 82%.

Noutro trabalho também já apresentado, o de Alkhatib, El Barachi e Shaalan (2019), a classificação é feita considerando as dimensões do relatório de incidente a ser gerado, desde tipo de evento, data, localização até impacto e escopo do evento. Os autores utilizaram cinco classificadores, dois dos quais aparencem na presente dissertação (*Naive Bayes* e SVM). É também feito o teste com e sem a técnica de *stemming* na parte de pré-processamento (aqui comentada no Passo 7). O classificador SVM conseguiu o 2º melhor desempenho, em ambos os casos,

considerando, também, o tempo de resposta do processamento da classificação, afinal, a proposições daquele trablaho era a de fornecer relatórios em tempo real.

Está-se trazendo tais resultados, da literatura revisada, justamente por se entender que cada situação específica pode acabar resultando em desempenhos diferentes, para os modelos que aqui são propostos, de modo que se está optando por treinar quatro modelos distintos em função da variabilidade que pode haver no desempenho da classificação.

Já referente, mais especificamente, à Análise de Sentimento seguindo a lógica não supervisionada, do **passo 20**, utilizar-se-á o método *Valence Aware Dictionary and sEntiment Reasoner* (*VADER)*[29] *Sentiment Analysis*, que é um método baseado em léxico, próprio para opiniões expressas em Redes Sociais. Tal estratégia foi utilizada por Abdul-Rahman et al., (2021), obtendo o melhor desempenho frente a outros classificadores supervisionados, no trabalho de Kazmaier e Vuuren (2020).

É feita, entretanto, desde já, a ressalva, contudo, que aqui é feito a pressuposição de que seu desempenho pode ser comprometido visto se tratar, neste caso, de trabalho que levará em consideração conteúdo textual em português.

Correlata à Análise de Sentimento, tem-se a Análise de Emoções, a qual tem por fito identificar *labels* de emoções, como raiva, felicidade, no conjunto de dados, tal qual fizeram Jain e Kumar (2018), Alkhatib et al., (2020), Liu, Teng e Gong, (2021) e Adamu et al., (2021). Utilizar-se-á de um widget do Orange denominado *Twitter Profiler*, pelo método de *Ekman*, tal qual Jain e Kumar (2018).

É buscado, através da Análise de Emoções, a abstração de um significado mais aprofundado, para além da polaridade de sentimento, tentando, assim, identificar a categoria de sentimento vivenciado pelos *netizens*. A Análise de Emoções é representada no **passo 21**.

Na esteira do que defende a literatura, acerca de, em detrimento de uma Análise de Sentimento “seca”, conjugar outros tipos de análise, é que define-se, nos **passos 22 e 23**, outras estratégias analíticas (WANG et al., 2015, ANDREWS et al., 2016, ALIZADEH; SARLAR; BIRGPUME, 2019, GASCO et al., 2019, EL-DIRABY; SHALABY; HOSSEINI, 2019, ALKHATIB; EL BARACHI; SHAALAN, 2019, ALKHATIB et al., 2020, ADAMU et al., 2021, MARZOUKI et al., 2021), quais sejam, neste caso, a análise descritiva do sentimento, representada pela visualização das relações das

[29] Acesso em 5 de Fevereiro de 2022. Disponível em: https://github.com/cjhutto/vaderSentiment.

variáveis e sua influência no Sentimento, observada por meio de uma *Decision Tree*, assim como a análise da frequência de palavras no *Dataset* limpo, para cada categoria de sentimento, a ser visualizada através de uma *Word Cloud*, a exemplo de trabalhos como os de Joseph et al., (2017), Kazmaier e Vuuren (2020) e Adamu et al., (2021).

Aproveitando o ensejo, passa-se, na sequência à breve seção Apresentação, desta Macro-Etapa 2.

3.2.5 Apresentação_2

A presente seção tem por finalidade a apresentação dos meios de visualização escolhidos, de modo a serem explicitados os resultados das estratégias de análise trazidas na seção anterior, Entendimento_2.

A existência da presente seção, para além de ser uma das etapas delimitadas na *CUP Framework,* corresponde também à resposta a um anseio/lacuna trazida por Kazmaier e Vuuren (2020), onde os autores comentam que o *feedback* analítico, ou seja, a demonstração das análises feitas, era uma parte deficiente, pouco atendida, na literatura correlata.

Assim, do **passo 24 ao 31**, estão evidenciadas as escolhas, do modelo aqui proposto, para a visualização das análises.

No **passo 24** é proposta a visualização das relações das variáveis com uma *Decision Tree.* O que se pretende, com este passo, é a visualização, na prática, de uma análise multivariada, que leve em conta vários metadados extraídos junto dos *tweets,* referente ao próprio *tweet* (quantidade de *retweets*, por exemplo) ou concernentes ao usuário que produziu o *tweet* (número de seguidores, por exemplo), e utilizá-los como determinantes (*features*) para a variável *meta* Sentimento.

A visualização por meio do diagrama de árvores permitirá aferir qual *feature* (metadado numérico) é o mais influente para o resultado da variável categórica Sentimento, para cada uma de suas dimensões, quais sejam, os próprios sentimentos indicados: positivo, negativo ou neutro. É também possível verificar as subrrelações presentes, entre a variável com maior influência no sentimento e as demais *features* numéricas.

Já no que diz respeito **ao passo 25**, intenta-se visualizar a evolução do sentimento de maneira temporal, comparando as datas em que houve o maior volume em produção de *tweets* com a variação do sentimento do *Dataset*. A inclusão da

temporalidade, nas análises, também é recomendação da literatura (ANDREWS et al., 2016, ALSAEDI; BURNAP; RANA, 2017, ALIZADEH; SARLAR; BIRGPUME, 2019, ALKHATIB et al., 2020, MARZOUKI et al., 2021).

Com relação **ao passo 26**, que trata da visualização das palavras mais frequentes no *Dataset* por meio de uma *Word-Cloud,* justifica-se que tal passo permita a verificação exploratória dos termos mais frequentes dos *tweets* componentes do conjunto de dados, no período estudado, de modo a poder alimentar outras extrações de dados temáticas no futuro, tratando de elementos latentes ao tópico que se mirou, considerando, ainda, diferentes palavras para cada categoria de sentimento. Ainda, a utilização de *Word-Clouds* é prática adotada em trabalhos da literatura (JOSEPH et al., 2017, KAZMAIER E VUUREN, 2020, e ADAMU et al., 2021).

Para a visualização da *Emotion Analysis*, tal qual discriminado no **passo 21**, se opta pela visualização por meio de *Box-Plot*, ou diagrama de caixa, representado, aqui, pelo **passo 27**. Tal método de visualização, para o que se intenta, permite observar predominância de uma emoção frente a outras, de maneira fácil de visualizar.

A visualização trazida no **passo 28** diz respeito à observação da distribuição dos sentimentos – avaliados tanto pela estratégia Não-Supervisionada como pela Supervisionada – por meio de um *Heat map*, ou mapa de calor. Verifica-se que tal abordagem é satisfatória, em geral, para métodos que envolvem clusterização de dados. Neste caso, por meio de um degrade de cores, é possível visualizar os dados (no caso, *tweets*) reputados como mais similares, havendo, ainda, a opção de "transformação" do mapa de calor em um dendograma, permitindo a observação dos níveis de similaridade entre os dados analisados.

Passando aos métodos de visualização escolhidos para observar os resultados da estratégia supervisionada de classificação dos sentimentos (**passo 19**), tem-se os passos representados de 29 a 31.

Referente ao **passo 29**, sua utilização se justifica na medida em que uma *Confusion Matrix* é apta a permitir a visualização do desempenho dos modelos treinados, revelando sua acurácia e proporção de classificações errôneas de sentimentos, como falsos positivos e falsos negativos. A referida estratégia é complementar à visualização dos desempenhos dos modelos treinados, conforme trazidos pela tabela do *Widget Test & Score* (**passo 30**), do *Orange Data Mining*, onde o desempenho dos modelos é comparado segundo cinco métricas: AUC, CA, F1,

Precision e *Recall.* Aproveitando o ensejo, estas cinco métricas indicadoras do desempenho dos quatro modelos da estratégia supervisionada, do passo 19.1.1 ao 19.1.4, são explicadas abaixo:

- AUC: Tal acrônimo traduz-se enquanto área debaixo da curva (*Area Under the Curve,* no inglês) e mede a performance geral de um modelo de classificação, sendo interpretado enquanto maior melhor, podendo seu valor variar entre 0 e 1 (JAMES et al., 2013)
- CA[30]: Do inglês, *Classification Accuracy*, que em tradução livre, indica a acurácia do modelo.
- F1[31]: A medida F1-*Score,* é dada pelo cálculo da média harmônica entre a *Precision* e o *Recall,* ambos indicadores abaixo, também sendo interpretado quanto maior, melhor.
- *Precision*[32]: Em tradução livre, a Precisão é a razão entre tp/(tp+fp), onde tp é entendido como *true positives*, ou os verdadeiros positivos, e fp são os *false positives,* ou falso positivos, também sendo interpretado como quanto maior, melhor.
- *Recall*[33]: O *Recall* é calculado como a razão entre tp/(tp+fn), onde tp consta explicado acima, e fn traduz-se enquanto falso negativos (*false negatives*), havendo, novamente, a interpretação de quanto maior, melhor.

Finalmente, o que se pretende, com o **passo 31**, é observar, de modo macro, a proporção de sentimentos observada no *Dataset* avaliado. Tal visualização tem o intuito de oferecer um panorama geral e inicial, para a análise, sendo feita a opção, ainda, de se apresentar a visualização proposta aqui, por primeiro, ao usuário, antes de adentrar às outras estratégias de visualização, das demais análises.

Tendo sido vencida a seção da metodologia, do presente trabalho, passa-se, na sequência, a análise de discussão dos resultados.

---

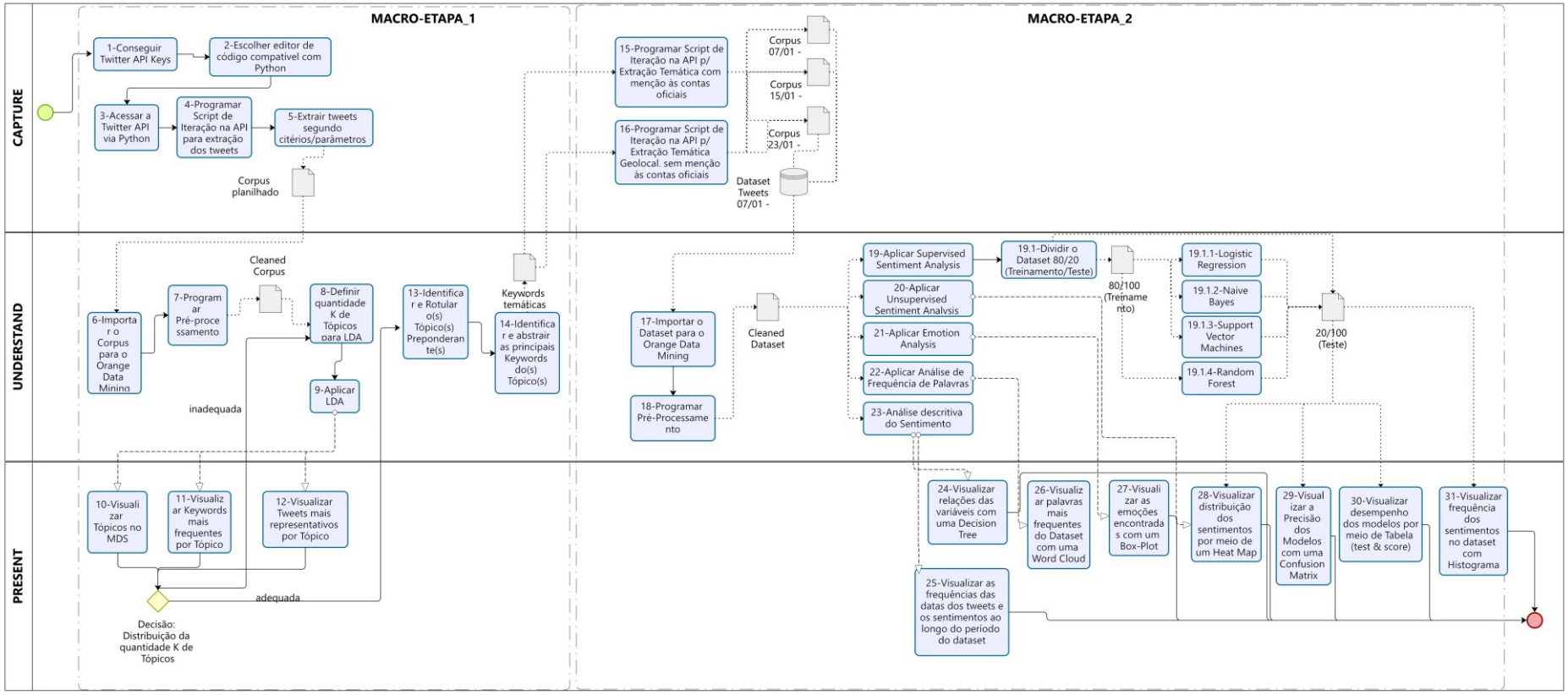
[30] Acesso em 8 de Fevereiro de 2022. Disponível em: https://orange3.readthedocs.io/projects/orange-data-mining-library/en/latest/reference/evaluation.cd.html.

4 RESULTADOS

4.1 Apresentação do Modelo Inicial

É apresentado, abaixo, na Figura 8, o fluxograma do modelo, cujo modelamento se deu com auxílio do *software* gratuito *Bizagi Modeler* – construído a partir das *Frameworks* bases, com esteio nos passos explicados e fundamentados na seção de construção do modelo, trazida no capítulo de Metodologia.

Figura 8 - *Framework* proposta (*CUP/ECCO*)



Fonte: dados de pesquisa (2022)

A aplicação num caso prático, para teste do modelo, é feita na seção abaixo, onde se pode verificar sua operacionalidade, tendo por local de teste a cidade paranaense de Maringá, em dois testes de modelo. Antes, porém, são discriminadas, desde logo, algumas escolhas/opções tomadas, referente à operacionalização e etapas do modelo.

Referente ao **passo 2**, fora escolhido o editor *PyCharm Community Edition*, um ambiente de programação integrado, gratuito, com base em *open source*[34] que fornece verificações inteligentes e dicas para o código sendo escrito, dentre outras funcionalidades, que são úteis à pessoas com conhecimento básicos de programação.

Referente ao **passo 4** os pacotes previamente comentados, e utilizados, no caso prático foram o *Tweepy*[35], que é uma biblioteca *Python* de fácil usabilidade para acesso à *Twitter* API, e o *Pandas*[36], através do qual, procedeu-se à aposição dos dados extraídos a uma planilha no formato .csv, facilitando a análise. Ambos os pacotes foram instalados por meio do instalador de pacotes *pip*[37].

Referente às contas dos usuários centrais escolhidas, de Maringá, no *Twitter,* seus *nicknames*/nomes de usuário são: @prefeiturademga e @UlissesMaia, as quais correspondem às contas da prefeitura e do prefeito de Maringá, respectivamente. As referidas contas possuíam, na data de 31/01/2022, o seguinte panorama[38] [39], conforme Figura 7:

---

[34] Acesso em 5 de Fevereiro de 2022. Disponível em: https://www.jetbrains.com/pt-br/pycharm/download/#section=windows.
[35] Acesso em 5 de Fevereiro de 2022. Disponível em: https://www.tweepy.org/.
[36] Acesso em 5 de Fevereiro de 2022. Disponível em: https://pandas.pydata.org/.
[37] Acesso em 5 de Fevereiro de 2022. Disponível em: https://pypi.org/project/pip/.
[38] Acesso em 31 de Janeiro de 2022. Disponível em: https://twitter.com/prefeiturademga?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor.
[39] Acesso em 31 de Janeiro de 2022. Disponível em: https://twitter.com/UlissesMaia?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor.

Figura 9 - Panorama das contas centrais

Fonte: dados de pesquisa (2022)

Referente ao **passo 16** foi utilizado na utilização do filtro *geocode*, foram inseridas as coordenadas geográficas -23.421031,-51.937609 e mantido o *range* estipulado de 10km de raio. Ainda referente ao **passo 16,** foram identificadas cinco contas predominantes, em volume de *tweets* produzidos, de cunho meramente informativo, que correspondiam às referidas alcunhas: @cbnmaringa, @Maringapost, @diario_maringa, @MaringaCom e @portalgmconline. Assim, também foi setado filtro de modo a não coletar *tweets* produzidos por estas citadas contas, por não refletirem, em teoria, opiniões diretas dos *netizens* de Maringá.

4.2 Aplicação do Modelo em Caso prático

A presente seção secundária está subdividida em três partes: uma para apresentação das características da cidade *lócus* de aplicação do modelo, e outras duas com os testes do modelo, com os respectivos resultados, de fato.

4.2.1 Apresentação do *Lócus* - Maringá

Tendo, o presente trabalho, por *lócus,* o contexto das cidades, ratifica-se, uma vez mais, a escolha da cidade paranaense de Maringá, enquanto objeto deste estudo,

pelas características trazidas nesta seção, justificando-se, assim, o *lócus* do estudo, conforme constante no objetivo geral.

Maringá é uma cidade situada no norte do estado do Paraná, de porte médio (RODRIGUES, 2004, SAVI; CORDOVIL, 2015), cuja população está estimada em 436.472 pessoas[40]. Teve sua colonização planejada pela Companhia de Terras Norte do Paraná, e sua fundação oficial em 10 de maio de 1947[41].

Maringá possui uma área territorial de 487,012 km², fazendo divisa[42] com 8 municípios, quais sejam Ângulo, Astorga, Iguaraçu, Floresta, Marialva, Sarandi, Paiçandu e Mandaguaçu.

O executivo municipal é comandado pelo prefeito Ulisses De Jesus Maia Kotsifas, atualmente em seu segundo mandato, tendo iniciado em 2017 e tendo sido reeleito para a gestão iniciada em 2021[43].

Com um resultado de 0,808, Maringá ocupa a posição de número 23, a nível nacional, no Índice de Desenvolvimento Humano (IDH)[44], tendo, ainda, um Produto Interno Bruto (PIB) *per capita*[45] de 44.442,52 reais. Maringá apresenta ainda um Índice Firjan de Desenvolvimento Municipal (IFDM) – responsável pela medição do desempenho de municípios brasileiros considerando três searas, quais sejam Emprego e Renda, Educação e Saúde – de 0.8646[46] o 5º melhor, entre as cidades paranaenses, e o 29º melhor do Brasil.

Referente à sua classificação territorial, é categorizada enquanto Capital Regional B, na hierarquia urbana, dada a cidades que são “centros de referência no interior dos estados”, sendo, as Capitais Regionais, entendidas como centros urbanos com alta concentração de atividades de gestão, mas com alcance menor em termos

---

[40] Acesso em 27 de Janeiro de 2022. Disponível em: https://cidades.ibge.gov.br/brasil/pr/maringa/panorama.
[41] Acesso em 27 de Janeiro de 2022. Disponível em: http://www2.maringa.pr.gov.br/turismo/?cod=nossa-cidade/2.
[42] Acesso em 1 de Março de 2022. Disponível em: https://cidades.ibge.gov.br/brasil/pr/maringa/panorama.
[43] Acesso em 1 de Março de 2022. Disponível em: http://www2.maringa.pr.gov.br/turismo/?cod=prefeitos.
[44] Acesso em 1 de Março de 2022. Disponível em: https://www.br.undp.org/content/brazil/pt/home/idh0/rankings/idhm-municipios-2010.html.
[45] Acesso em 1 de Março de 2022. Disponível em: https://www.cidadessustentaveis.org.br/painel-cidade/detalhes/4121.
[46] Acesso em 7 de Março de 2022. Disponível em: https://www.firjan.com.br/ifdm/consulta-ao-indice/ifdm-indice-firjan-de-desenvolvimento-municipal-resultado.htm?UF=PR&IdCidade=411520&Indicador=1&Ano=2016.

de região de influência”[47]. Assim sendo, a cidade de Maringá não se confunde com uma metrópole, exercendo, no entanto, influência central na região onde está localizada, no interior do estado do Paraná, seja por seu volume demográfico ou oferta de estruturas de bens e serviços.

Reforça-se, ainda, a posição de Maringá enquanto a 20ª cidade mais inteligente do Brasil de acordo com último *ranking Connected Smart Cities* (era a 25ª cidade mais inteligente do Brasil, no *ranking* anterior), sendo ainda a cidade mais inteligente do estado do Paraná[48] [49], entre aquelas com população entre 100 mil e 500 mil habitantes, e a sétima, a nível nacional (era a nona, neste critério, no *ranking* anterior), conforme este critério populacional, segundo o *ranking Connected Smart Cities*. Tendo em vista, ainda, as 20 primeiras cidades classificadas no *ranking* citado*,* dentro do estrato que compreende cidades com populações entre 100 e 500 mil habitantes, verificou-se que 13 destas possuem, assim como Maringá, a classificação, pelo IBGE, enquanto Capital Regional, representando, assim, potenciais *lócus* de estudo.

Utilizou-se como ponto de partida o *ranking Connected Smart Cities,* para identificação do *lócus* de estudo, cuja sistemática de mensuração da inteligência das cidades leva em conta 11 eixos temáticos, compostos, no total, por 75 indicadores. Os eixos em questão são: Mobilidade, urbanismo, Meio Ambiente, Tecnologia e Inovação, Empreendedorismo, Educação, Saúde, Segurança, Energia, Governança e Economia[50], os quais podem ser situados dentro das 6 dimensões de Cidade Inteligente, propostas por Giffinger et al., (2007).

Tendo sido feita a explanação da escolha da cidade *lócus* de teste do modelo, parte-se, na sequência, aos resultados fáticos, encontrados.

4.2.2 Resultados do Primeiro Teste do Modelo

[47] Acesso em 27 de Janeiro de 2022. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101728_folder.pdf.
[48] Acesso em 28 de Fevereiro de 2022. Disponível em: http://www2.maringa.pr.gov.br/site/noticias/2021/09/02/ranking-aponta-maringa-como-a-cidade-mais-inteligente-do-pr-com-populacao-entre-100-mil-e-500-mil-habitantes/38331.
[49] Acesso em 12 de janeiro de 2023. Disponível em: http://www.maringa.pr.gov.br/site/noticias/2022/11/01/maringa-e-a-7-cidade-mais-inteligente-do-brasil-entre-os-municipios-de-100-mil-a-500-mil-habitantes-aponta-ranking/40605
[50] Acesso em 12 de janeiro de 2023. Disponível em: https://ranking.connectedsmartcities.com.br/.

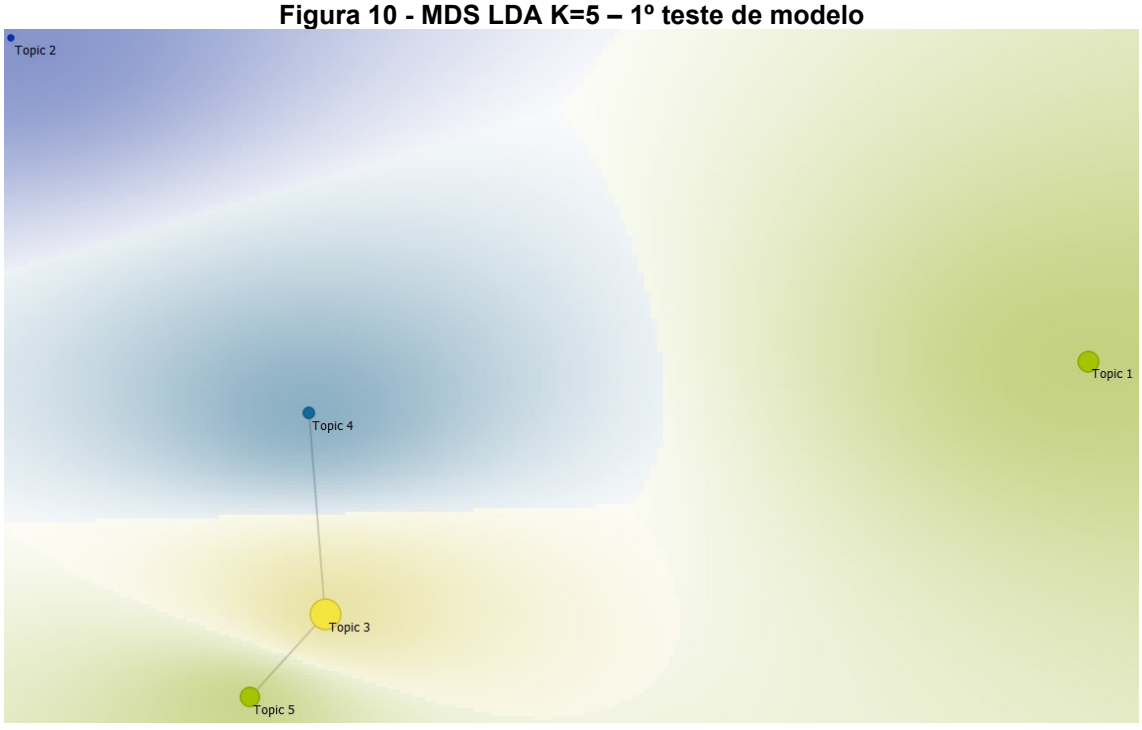
Inicia-se, desde logo, a discussão de resultados, tratando de apresentar o *Corpus* planilhado de documentos que serviu de *input* à primeira etapa de análise (Entendimento_1).

Foram reunidos, 429 *tweets* únicos, não repetidos – excluídos *retweets* – no período compreendido entre os dias 03 e 12 de janeiro de 2022. Conforme já comentado, tais *tweets* foram extraídos seguindo os parâmetros setados discutidos **no passo 5**, do fluxo metodológico proposto.

Passando pelos passos 6 e 7, conforme discriminado no capítulo da metodologia, chega-se ao **passo 8**, de definição da quantidade K, de tópicos, de modo a rodar a LDA. Por padrão, foi definido, inicialmente, K=10, e, na visualização iterativa conjugada com o MDS (**passo 10**), verificou-se *clusters* de tópicos cuja proximidade indicavam se tratar de assuntos similares ou iguais.

Assim, rodou-se novamente a LDA, com K=5. O resultado, foi conforme pode ser visto abaixo, na Figura 10:

Figura 10 - MDS LDA K=5 – 1º teste de modelo

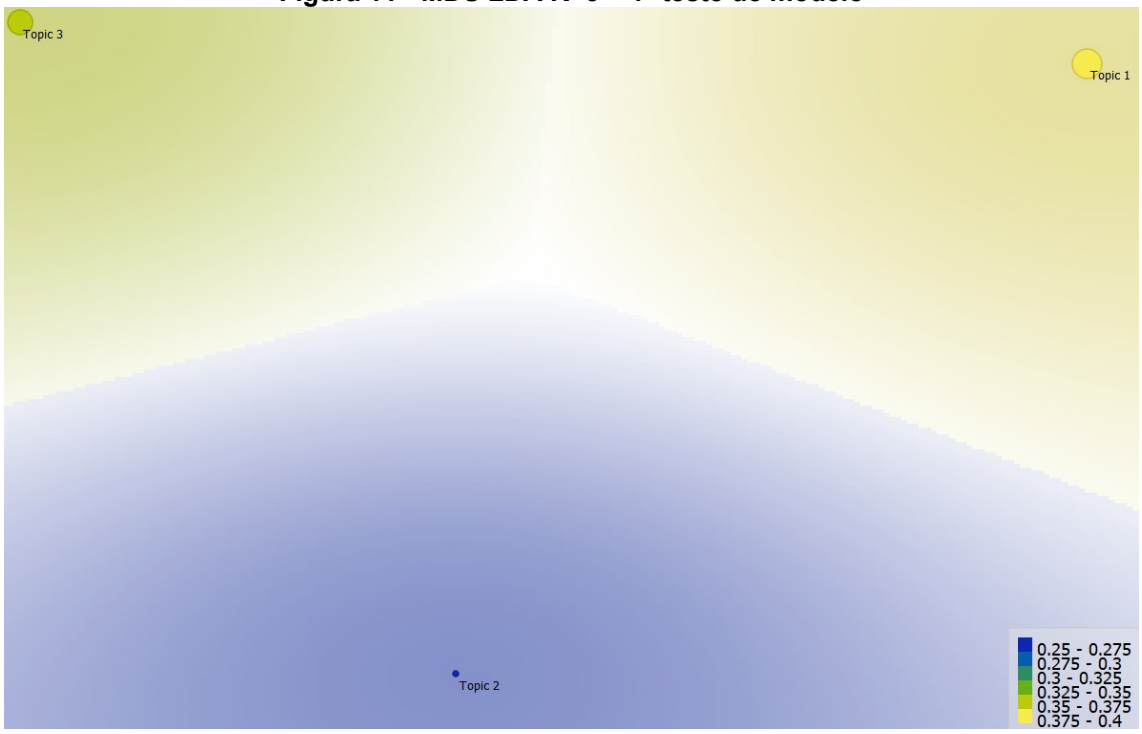


Fonte: dados de pesquisa (2022)

Pela distribuição visualizada no MDS, pode-se perceber, que os três tópicos mais ao centro da figura têm similaridade de assuntos abordados, reforçada, inclusive, pela opção setada, no *Orange*, de indicação de similaridade entre pares (*Show similar pairs*).

Assim, na prática, interpretou-se haver, na realidade, três tópicos definidos, a julgar pela distribuição observável via MDS, sendo então, ao final, setados K=3 tópicos para aplicação da LDA, a qual resultou na seguinte distribuição, tal qual Figura 11:

Figura 11 - MDS LDA K=3 – 1º teste de modelo



Fonte: dados de pesquisa (2022)

Percebe-se, tão logo, a predominância dos tópicos 1 e 3, no que diz respeito à sua presença no *Corpus* de documentos analisados.

Pela leitura dos *tweets* mais representativos de cada tópico (**passo 12**) e suas *Keywords* componentes, a rotulação (**passo 13**) feita foi conforme abaixo:

- TÓPICO 1 – Vacinação; Uso de Máscara; Economia; Setor de eventos.
- TÓPICO 2 – Infraestrutura; Orçamento; Vacinação
- TÓPICO 3 – Vacinação; Testagem; Atendimentos

O que se pode perceber, é que assuntos relacionados à pandemia como vacinação acabaram presentes nos três tópicos, sendo difícil fazer sua dissociação.

Assim, não obstante outras temáticas tivessem sido identificados, como a preocupação dos *netizens* com a Economia, Orçamento (e aplicação dos repasses do Governo Federal) e Infraestrutura, o tema que perpassou a maior parte das discussões foi relacionado à Pandemia, Vacinação e afins. Mesmo a questão

referente ao Setor de eventos, ao que se pode analisar, acabava fazendo referência à pandemia e as medidas de restrição associadas.

Assim, por entender que o assunto preponderante observado, sobretudo nos tópicos 1 e 3 dizia respeito à Saúde, de um modo geral, mas mais especificamente, à Pandemia e à Vacinação, o que se fez foi formar uma lista (**passo 14**) de *Keywords*, por meio das palavras significativas desta temática, e utilizá-la como insumo para iterações na *Twitter* API, de modo à extração de dados referentes à esta matéria.

As palavras significativas elencadas são as que constam discriminadas na sequência:

Quadro 8 - *Keywords* temáticas Saúde

saude OR covid OR covid-19 OR UPA OR máscara OR máscaras OR isolamento OR isolado OR médico OR médicos OR infectologista OR infectologistas OR teste OR testes OR pandemia OR pandemico OR vacinação OR vacina OR vacinado OR vacinados OR vacinar OR posto OR postinho OR passaporte OR reforço

Fonte: dados de pesquisa (2022)

De posse desta lista, o que se fez na sequência foi a programação de dois *scripts* de iteração na *Twitter* API, para extração dos *tweets* temáticos (conforme **passos 15 e 16**).

Na Lógica de extração do **passo 16** foram identificadas, preliminarmente, cinco contas predominantes, em volume de *tweets* produzidos, de cunho meramente informativo, que correspondiam às referidas alcunhas: @cbnmaringa, @Maringapost, @diario_maringa, @MaringaCom e @portalgmconline. Assim, também foi setado filtro de modo a não coletar *tweets* produzidos por estas citadas contas, por não refletirem, em teoria, opiniões diretas dos *netizens* de Maringá.

Os *scripts* foram rodados separadamente, por três vezes, de modo a contornar as limitações temporais da *Twitter* API, e conseguir resultados que considerassem um período maior do que 9 dias, apenas.

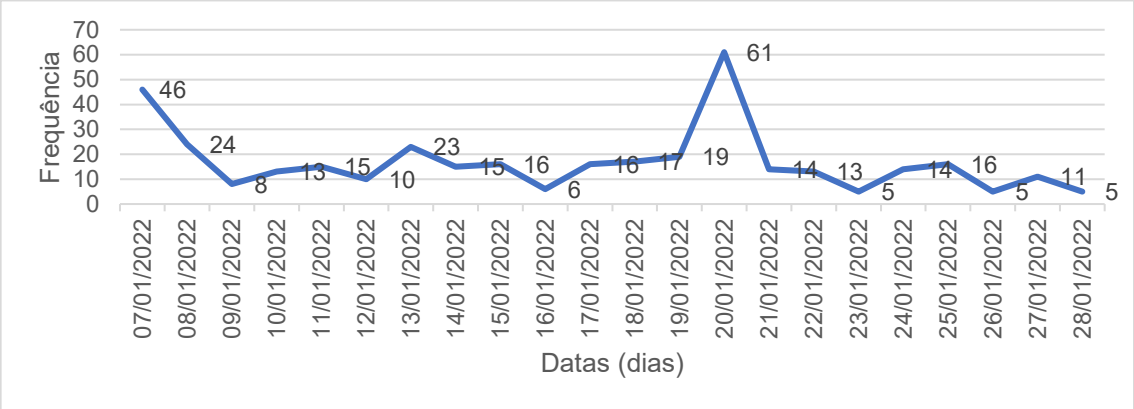
Foram rodados os *scripts* para extração dos *tweets* temáticos compreendendo o período das seguintes datas:

- 07/01 – 15/01;
- 15/01 – 23/01;
- 23/01 – 28/01.

Os *Corpus* de documentos de cada uma das datas foram juntados, onde procedeu-se à exclusão de *tweets* repetidos. Restou, então, um *Dataset* com apenas

tweets únicos, temáticos, referenciados em Maringá, resultando num total de 372 documentos, distribuídos num intervalo de 21 dias, conforme gráfico de linhas, abaixo, ilustrado no Gráfico 1.

Gráfico 1 - Frequência diária de tweets no Dataset – 1º teste de modelo



Fonte: dados de pesquisa (2022)

A média de postagem diária, no período foi de 16,9 *tweets*, tendo, esta amostra, um desvio padrão de 13,17. Assim, deste conjunto de dados, os dias em que se observou uma maior disparidade na frequência de postagens, considerando o a média, foram nos dias de 07/01 e 20/01, com 46 e 61 *tweets* diários, respectivamente.

O pico de frequência referente à 07/01 acredita ser devido em função de, na referida data, ter sido publicado, pelo executivo municipal, o Decreto 19/2022[51], que, dentre outras disposições, previa a manutenção do uso obrigatório de máscara para todas as pessoas acima de 3 (três) anos, inclusive em espaços abertos ao público, além da previsão de multa no valor entre R$ 150,00 (cento e cinquenta reais) e R$ 550,00 (quinhentos e cinquenta reais) para os que descumprirem a normativa.

Já o pico de 20/01, acredita-se ser a dois fatores. O primeiro deles, devido ao Decreto 86/2022[52], publicado no dia anterior, ou seja, 19/01, prevendo restrições à realização de eventos, tal como a exigência da apresentação de comprovante de vacinação e aplicação de multa em caso de descumprimento. O segundo deles, muito em função do primeiro fator, foi a suspensão das aulas presenciais pela Universidade Estadual de Maringá (UEM), por meio da Portaria 029/2022[53].

---

[51] Acesso em 9 de Fevereiro de 2022. Disponível em: http://www2.maringa.pr.gov.br/cdn-imprensa/DECRETO19-2022.pdf.

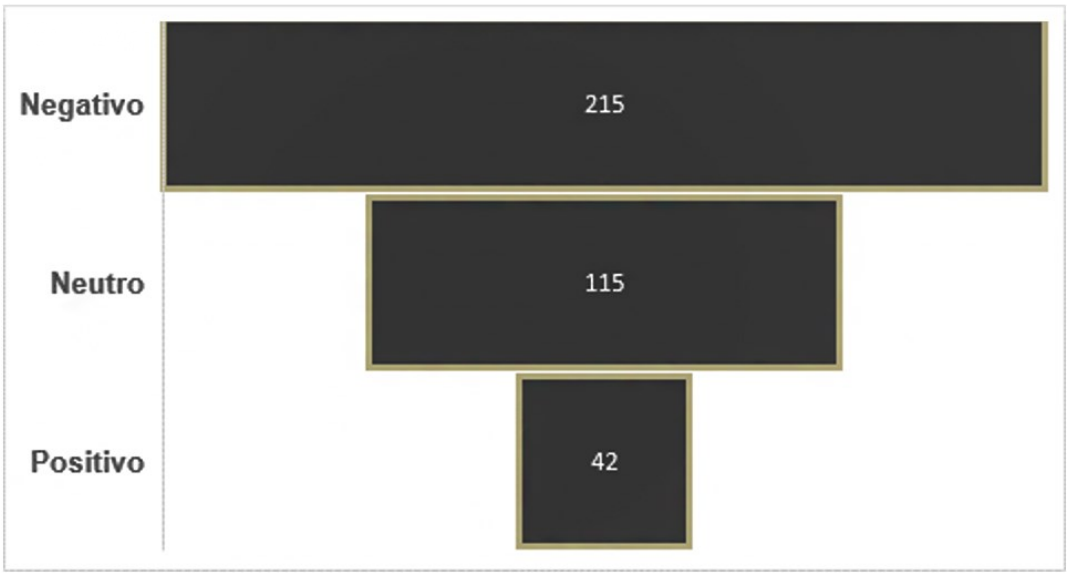
[52] Acesso em 9 de Fevereiro de 2022. Disponível em: http://www2.maringa.pr.gov.br/sistema/arquivos/9f26c0431561.pdf.

[53] Acesso em 9 de Fevereiro de 2022. Disponível em: http://noticias.uem.br/images/2021/Portaria_Suspensa%CC%83o_Presencial.pdf.

Procedeu-se, então, já tendo em mente às ações de classificação supervisionada do sentimento, representadas pelo **passo 19**, à marcação manual do sentimento de cada um dos 379 *tweets* do *Dataset*, em positivo, negativo ou neutro.

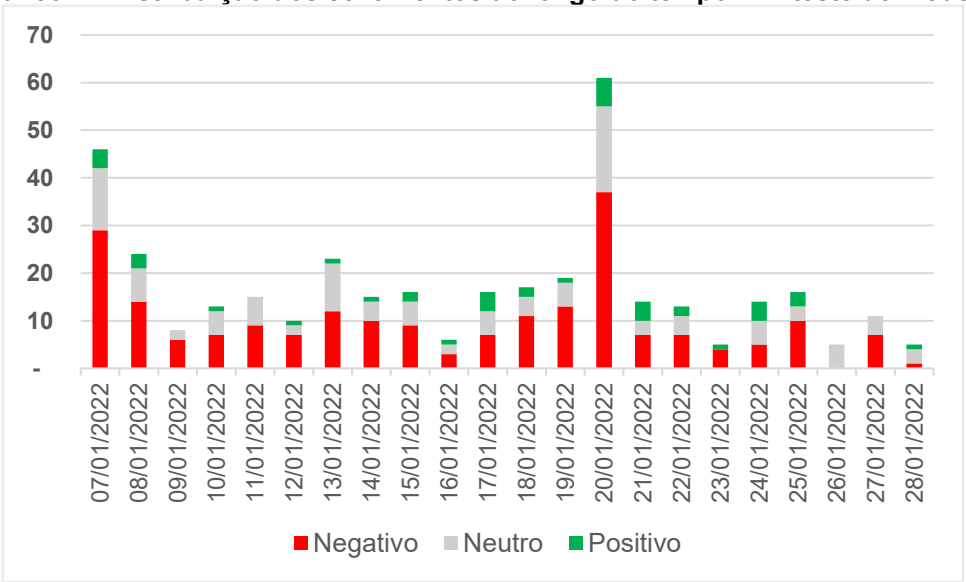
Desde logo são apresentados, tendo em vista os **passos 25 e 31**, a distribuição da frequência dos sentimentos verificados ao longo do *Dataset*, bem como sua distribuição ao longo do período considerado para extração dos dados. Gráficos 2 e 3, respectivamente:

Figura 12 - Frequência de Sentimentos no Dataset – 1º teste de modelo



Fonte: dados de pesquisa (2022)

Gráfico 2 - Distribuição dos sentimentos ao longo do tempo – 1º teste de modelo



Fonte: dados de pesquisa (2022)

Observa-se, de antemão, que o sentimento mais frequente observado, na classificação da polaridade, manualmente, foram *tweets* negativos. Isso condiz com o que foi encontrado na literatura sobre a temática, quando se afirma que geralmente os *netizens* utilizam o *twitter* para compartilhar suas frustrações, reclamações e afins (EL-DIRABY; SHALABY; HOSSEINI, 2019).

Tal predominância da polaridade positiva, de sentimento, foi observada ao longo de todo o período avaliado, mais facilmente visualizada nos picos dos dias 07/01 e 20/01.

Tratando agora acerca dos desempenhos dos quatro modelos da estratégia supervisionada, seguiu-se a diretriz de iteração dentro da etapa de entendimento dos dados, onde foram testados diferentes parâmetros – conforme sugerido por, visualizados no **passo 30**, nas tabelas de *test & score,* conforme sugerido por Kazmaier e Vuuren (2020).

Foram definidos o tamanho do *set* de treinamento em 95%, repetindo o teste 50 vezes. A título de comparação, foi definida uma rodada com a etapa de pré-processamento do *Corpus* de treinamento, e outra sem pré-processamento, ambas com abordagem lexical, operacionalizadas por meio do *widget Orange* de *bag of words*.

Os resultados são os que seguem, primeiro para o desempenho com pré-processamento (mas sem remoção das interrogações (?) e exclamações (!), conforme já explicado no capítulo da Metodologia), e outra sem pré-processamento:

Com pré-processamento, conforme Tabela 1:

Tabela 1 - Desempenho com pré-processamento – 1º teste de modelo

Classificador	AUC	CA	F1	Precision	Recall
SVM	0.826	0.6469	0.626	0.634	0.649
Random Forest	0.724	0.608	0.553	0.511	0.608
Naive Bayes	0.654	0.324	0.232	0.196	0.324
Logistic Regression	0.500	0.554	0.395	0.307	0.554

Fonte: dados de pesquisa (2022)

Sem pré-processamento, conforme Tabela 2:

Tabela 2 - Desempenho sem pré-processamento – 1º teste de modelo

Classificador	AUC	CA	F1	Precision	Recall
SVM	0.801	0.716	0.674	0.753	0.716

Random Forest	0.778	0.743	0.677	0.643	0.743
Naive Bayes	0.712	0.311	0.233	0.200	0.311
Logistic Regression	0.500	0.554	0.395	0.307	0.554

Fonte: dados de pesquisa (2022)

Interessante notar que o melhor resultado, foi obtido – segundo o índice AUC – aplicando o método de pré-processamento no conjunto de *tweets* para treinamento, para o modelo SVM, contudo, observou-se maior homogeneidade, entre os modelos, e desempenhos, em média, maiores, para todos os modelos, na rodada em que não se aplicou o pré-processamento.

De um modo geral, observou-se que os melhores modelos *foram Random Forest* e SVM.

Necessário, de todo modo, ter em mente que o desempenho para classificação de *tweets* negativos e neutros foi relativamente adequado, se levarmos em consideração os modelos SVM e *Random Forest*, contudo, para a classificação de *tweets* positivos, nenhum modelo teve muito sucesso, conforme pode ser observado na sequência, nos dois *sets*.

- Com pré-processamento:

a) Desempenho para classificação de *tweets* com sentimento Negativo, conforme Tabela 3:

Tabela 3 - Desempenho com pré-processamento para classificação de tweets negativos – 1º teste de modelo

Classificador	AUC	CA	F1	Precision	Recall
SVM	0.808	0.712	0.746	0.791	0.707
Random Forest	0.799	0.749	0.809	0.744	0.887
Naive Bayes	0.688	0.424	0.077	1.000	0.040
Logistic Regression	0.500	0.600	0.750	0.600	1.000

Fonte: dados de pesquisa (2022)

b) Desempenho para classificação de *tweets* com sentimento Neutro, conforme Tabela 4:

Tabela 4 - Desempenho com pré-processamento para classificação de tweets neutros – 1º teste de modelo

Classificador	AUC	CA	F1	Precision	Recall
SVM	0.812	0.713	0.632	0.552	0.740

Random Forest	0.796	0.765	0.621	0.673	0.576
Naive Bayes	0.779	0.657	0.570	0.490	0.680
Logistic Regression	0.500	0.667	0.000	0.000	0.000

Fonte: dados de pesquisa (2022)

c) Desempenho para classificação de *tweets* com sentimento Positivo, conforme Tabela 5:

Tabela 5 - Desempenho com pré-processamento para classificação de tweets positivos – 1º teste de modelo

Classificador	AUC	CA	F1	Precision	Recall
SVM	0.549	0.921	0.063	0.154	0.040
Random Forest	0.596	0.933	0.000	0.000	0.000
Naive Bayes	0.556	0.497	0.133	0.075	0.580
Logistic Regression	0.500	0.933	0.000	0.000	0.000

Fonte: dados de pesquisa (2022)

- Sem pré-processamento:

a) Desempenho para classificação de *tweets* com sentimento Negativo, conforme Tabela 6:

Tabela 6 - Desempenho sem pré-processamento para classificação de tweets negativos – 1º teste de modelo

Classificador	AUC	CA	F1	Precision	Recall
SVM	0.818	0.745	0.797	0.764	0.833
Random Forest	0.784	0.763	0.824	0.743	0.924
Naive Bayes	0.749	0.415	0.064	0.789	0.033
Logistic Regression	0.500	0.600	0.750	0.600	1.000

Fonte: dados de pesquisa (2022)

b) Desempenho para classificação de *tweets* com sentimento Neutro, conforme Tabela 7:

Tabela 7 - Desempenho sem pré-processamento para classificação de tweets neutros – 1º teste de modelo

Classificador	AUC	CA	F1	Precision	Recall
SVM	0.817	0.753	0.628	0.632	0.624
Random Forest	0.820	0.812	0.679	0.788	0.596
Naive Bayes	0.788	0.717	0.604	0.566	0.648

Logistic Regression	0.500	0.667	0.000	0.000	0.000

Fonte: dados de pesquisa (2022)

c) Desempenho para classificação de *tweets* com sentimento Positivo, conforme Tabela 8:

Tabela 8 - Desempenho sem pré-processamento para classificação de tweets positivos – 1º teste de modelo

Classificador	AUC	CA	F1	Precision	Recall
SVM	0.559	0.920	0.032	0.083	0.020
Random Forest	0.605	0.935	0.039	1.000	0.020
Naive Bayes	0.536	0.425	0.129	0.072	0.640
Logistic Regression	0.500	0.933	0.000	0.000	0.000

Fonte: dados de pesquisa (2022)

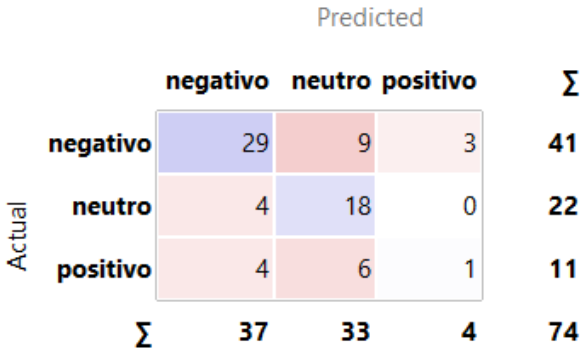
Acredita-se que essa dificuldade de classificação de *tweets* com sentimento Positivo seja devida, em parte, à complexidade da diferenciação de ironia e sarcasmo, de um elogio ou parabenização fática.

De modo a visualizar conjuntamente os resultados previstos, tanto de maneira correta quanto incorretamente, são apresentadas, abaixo, as *Confusion Matrixes*, para os métodos com e sem pré-processamento, dos modelos que performaram melhor nos testes conduzidos, quais sejam SVM e *Random Forest*:

- Com pré-processamento:

a) SVM, conforme Figura 20:

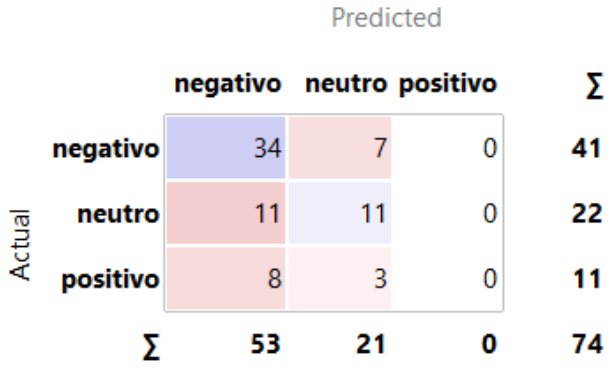
Figura 13 - Desempenho SVM com pré-processamento – 1º teste de modelo



Fonte: dados de pesquisa (2022)

b) Random Forest, conforme Figura 21*:*

Figura 14 - Desempenho *Random Forest* com pré-processamento – 1º teste de modelo

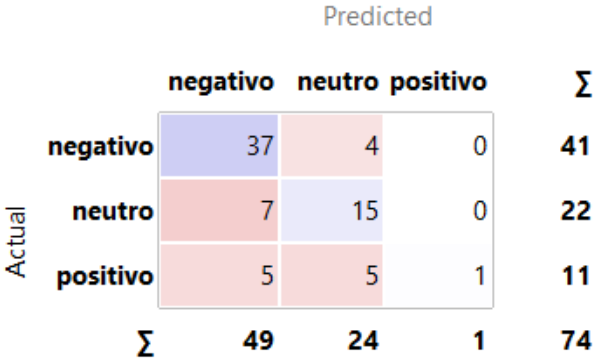


Fonte: dados de pesquisa (2022)

- Sem pré-processamento:

a) SVM, conforme Figura 22:

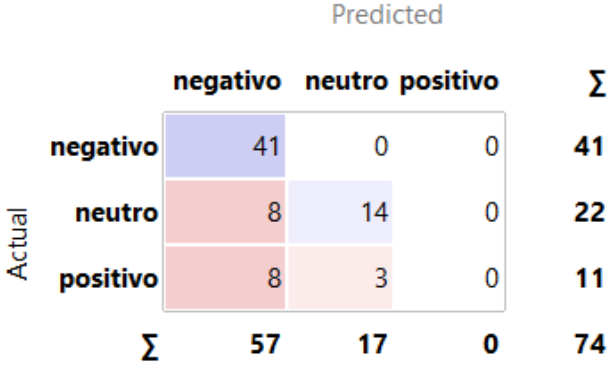
Figura 15 - Desempenho SVM sem pré-processamento – 1º teste de modelo



Fonte: dados de pesquisa (2022)

b) *Random Forest,* conforme Figura 23*:*

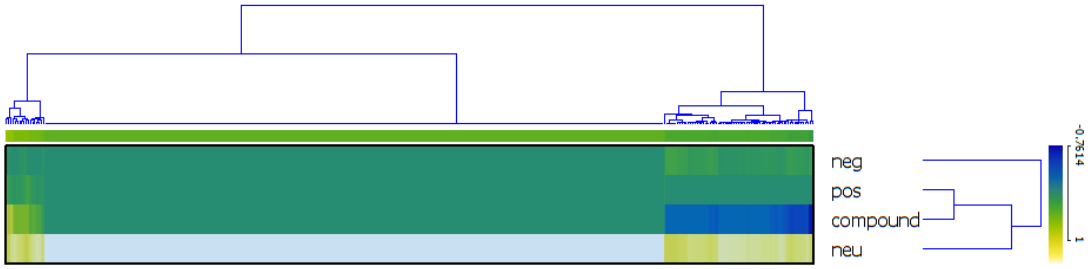
Figura 16 - Desempenho *Random Forest* sem pré-processamento – 1º teste de modelo



Fonte: dados de pesquisa (2022)

Referente à estratégia Não Supervisionada, para análise de sentimentos, foi feito como discriminado no **passo 20**. Por meio de um dendrograma, os sentimentos identificados foram clusterizados por proximidade, e constam conforme Figura 24:

Figura 17 - Dendograma de sentimentos análise não supervisionada



Fonte: dados de pesquisa (2022)

Pode-se observar, resolveu-se abstrair os três *tweets* mais significativos de cada polo (positivo e negativo), de modo a verificar, qualitativamente, o desempenho obtido.

Tweets positivos

@UlissesMaia De máscara já busquei minha marmita deliciosa do restaurante popular https://t.co/4nzGhOBSrM
Parabéns, hein @prefeiturademga! Cidade explodindo de casos de covid e o estádio WD cheio, até briga entre torcidas rolou! 😳 https://t.co/cvpxati2xi
prefs @UlissesMaia pelo amor de deus. Tô no upa zona sul desde as 10:30h e até agora não fui atendido. socorro

Tweets negativos

@prefeiturademga @UlissesMaia 11H PARA FAZER O TESTE RAPIDO DE COVID NO HOSPITAL MUNICIPAL eu não tenho nem palavras para falar o quanto eu tô indignada. Mas não vai ficar por isso mesmo, no mínimo uma reclamação formal no OuvidorSus será feita 😡😡😡😡😡 https://t.co/FMS7Grfpv7
po @UlissesMaia, vamos de UPA cheio nos próximos dias? nem parece que ta tendo 1k de casos por dia e o estadio Willie Davids lotadíssimo.
PREFEITO ULISSES MAIA - REPUBLICANOS DE MARINGÁ!? VOLTA A TOCAR O TERROR E COLOCA GCM NAS RUAS PARA MULTAR QUEM AO AR LIVRE TIVER SEM MÁSCARA E DECRETA "PASSAPORTE SANITÁRIO" SEM OUVIR A CÂMARA MUNICIPAL!? COM A PALAVRA AGORA A POPULAÇÃO E OS VEREADORES QUE A REPRESENTAM!? https://t.co/yvsDfN4dIe

Foram analisados mais *tweets* do que os mostrados aqui, contudo, pela pequena amostra trazida, já pode-se verificar que, pelo método não supervisionado,

foi possível a indicação, com alguma correção de *tweets* com sentimento negativo, mas não de *tweets* com sentimento positivo, sendo o primeiro aqui trazido, deste grupo, um exemplo de como *tweets* com ironia podem ser classificados erroneamente, corroborando com os resultados da abordagem supervisionada, trazida anteriormente.

No que diz respeito à estratégia de visualização do **passo 26**, tem-se as Nuvens de Palavras abaixo, visualizada separada por categoria de sentimento, seguindo a classificação manual feita (que serviu de *input* à análise de sentimento pela estratégia supervisionada). Houve uma etapa de pré-processamento prévio, para limpeza dos dados.

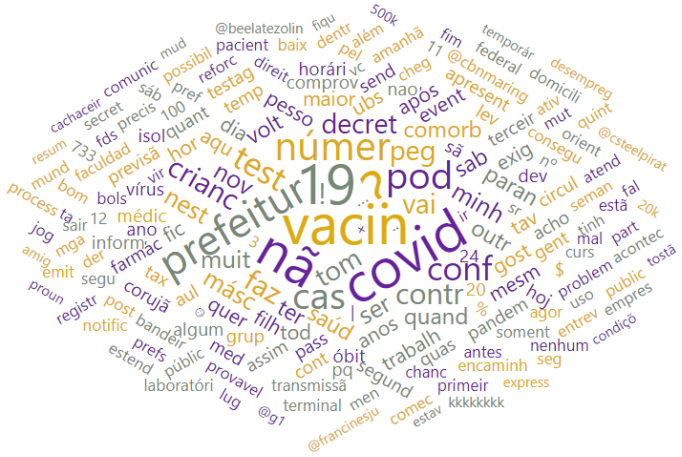
- *Word-Cloud* dos *tweets* classificados como Negativo, conforme Figura 25:

Figura 18 - *Word-Cloud* dos *tweets* classificados como negativo – 1º teste de modelo

Fonte: dados de pesquisa (2022)

- *Word-Cloud* dos *tweets* classificados como Neutro, conforme Figura 26:

Figura 19 - *Word-Cloud* dos *tweets* classificados como neutro – 1º teste de modelo



Fonte: dados de pesquisa (2022)

- *Word-Cloud* dos *tweets* classificados como Positivo, conforme Figura 27:

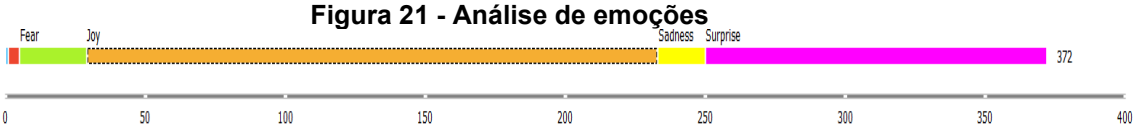
Figura 20 - *Word-Cloud* dos *tweets* classificados como positivo – 1º teste de modelo

Fonte: dados de pesquisa (2022)

Percebe-se a repetição de palavras, nas três categorias, as quais, não foram retiradas no pré-processamento para que não prejudicasse a interpretação das mais frequentes, tendo em vista a possível falta de contexto de palavras carregadas de significados positivos, ou negativos, e outras (temáticas) bastante frequentes.

Aplicando a análise de emoções do **passo 21**, e visualizando-as no **passo 27**, o resultado obtido é conforme abaixo, na Figura 20:

Figura 21 - Análise de emoções



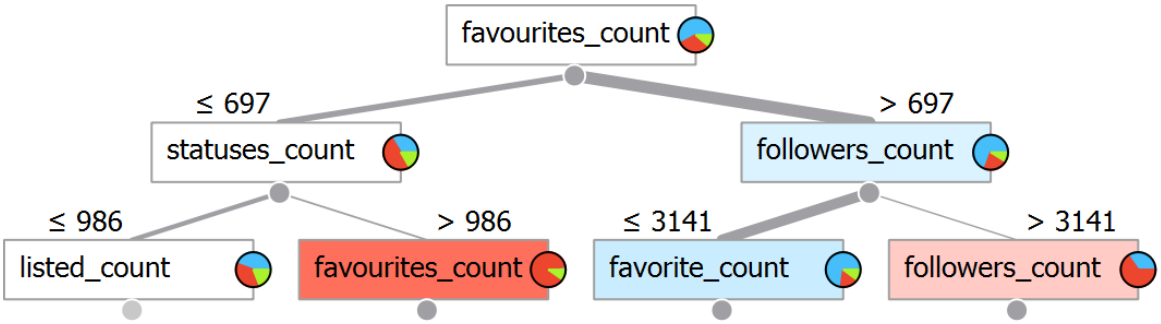
Fonte: dados de pesquisa (2022)

Desde logo vemos que houve uma má classificação, na medida em que "surpresa" e "felicidade" são os dois sentimentos que mais ocupam o *corpus*, segundo o processamento pelo *widget Tweet Profiler*, em detrimento de, na realidade, o *Dataset* ser composto de documentos na sua maioria com sentimento negativo.

A incorreta classificação pelo método de *Ekman*, seguindo uma abordagem *Unsupervised*, sugere que, neste caso, quem sabe uma abordagem mais adequada à identificação de emoções num *corpus* de documentos se dê através do treinamento de modelos, adotando uma abordagem supervisionada.

Com relação à visualização das relações das variáveis dos metadados dos *tweets*, tendo em vista o sentimento, tem-se o **passo 24**, o qual apresentou o seguinte panorama, conforme Figura 29:

Figura 22 - Diagrama de árvore relações entre metadados dos *tweets* – 1º teste de modelo



Fonte: dados de pesquisa (2022)

Discrimina-se, uma vez mais, o sentimento negativo como sendo o mais preponderante no *dataset,* sendo a polaridade identificada em 57,8% do total de *tweets*. Verificou-se que o metadado mais influente na polaridade do sentimento foi a *feature* denominada *"favourites_count"*. A *feature* em questão indica a contagem de "favoritações" ou "curtidas" que o usuário – produtor do *tweet* – deu, ao longo da vigência de sua conta.

Tendo em vista o resultado, foi calculada a média do referido metadado, para cada polaridade de sentimento, como pode se visto na Tabela 1, abaixo:

Tabela 9 - Média do metadado *favourites_count* por categoria de sentimento

CATEGORIA DE SENTIMENTO	Média do metadado "favourites_count"
negativo	19355,13
neutro	8848,88
positivo	6876,05

Fonte: dados de pesquisa (2022)

Observa-se que a maior média de “favoritações” históricas do usuário, ficou associado àqueles que produziram *tweets* com polaridade de sentimento negativa, o que pode indicar que usuários mais “ativos” do *Twitter* tendem a produzir *tweets* de cunho mais negativo, o que coaduna com a teoria de que o *Twitter* é a rede social onde o viés da produção do conteúdo está associado a externalização de reclamações, por parte dos seus usuários (EL-DIRABY; SHALABY; HOSSEINI, 2019).

Passa-se, na sequência, à apresentação dos resultados do segundo teste de aplicação do modelo.

4.2.3 Resultados do Segundo Teste do Modelo

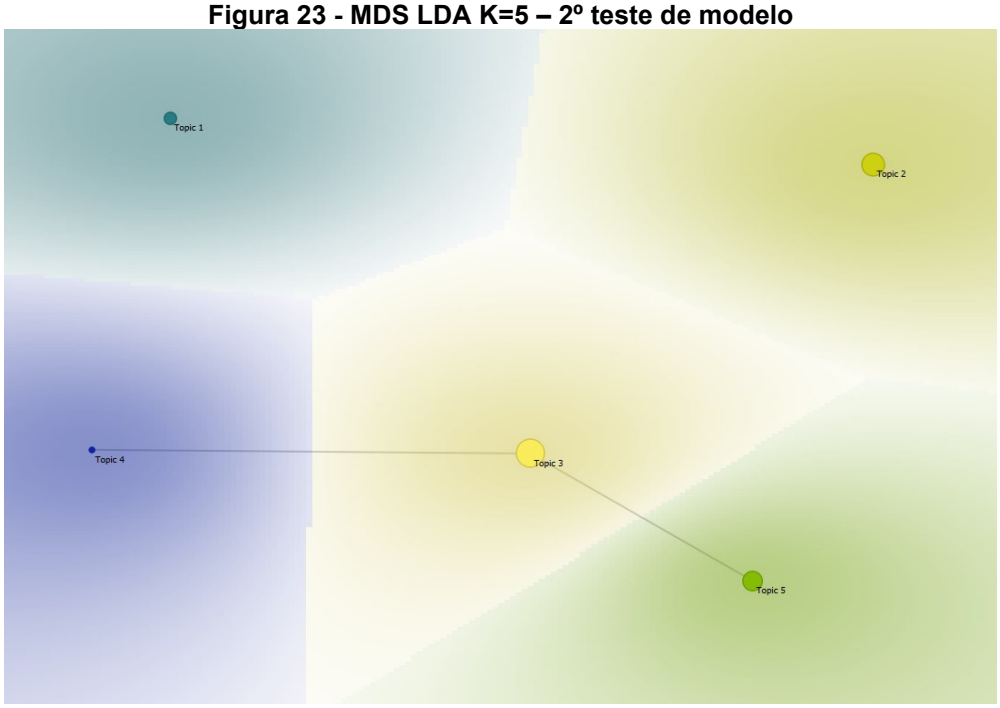
Desde logo deve ser comentado que algumas análises que foram feitas/propostas no primeiro teste de aplicação do modelo não foram repetidas neste segundo teste, por não terem sido frutíferas. As referidas análises não realizadas neste segundo teste são aquelas representadas pelos passos 20, 21, 24, 27 e 28, e dois dos classificadores da análise de sentimentos, propostos inicialmente por 19.1.1 e 19.1.2, não foram treinados, neste segundo teste. Intenta-se, também, deste modo, focalizar mais os propósitos do presente trabalho na modelagem de tópicos e análise de sentimentos.

Assim como na subseção anterior, inicia-se a discussão apresentando o *Corpus* planilhado de documentos que serviu de *input* à primeira etapa de análise (Entendimento_1), deste segundo tese de modelo.

Na rodagem do **passo 5** do modelo, foram reunidos, no período compreendido entre 23/07/2022 e 29/08/2022 um total de 626 *tweets* únicos, não repetidos, conforme parâmetros e opções já mencionados. Passando pelos passos 6 e 7, uma vez mais é definido, num primeiro momento K=10, de modo a rodar a LDA – **passos 8 e 9**, respectivamente – sendo utilizada, novamente, a visualização iterativa conjugada com o MDS, que é o **passo 10**.

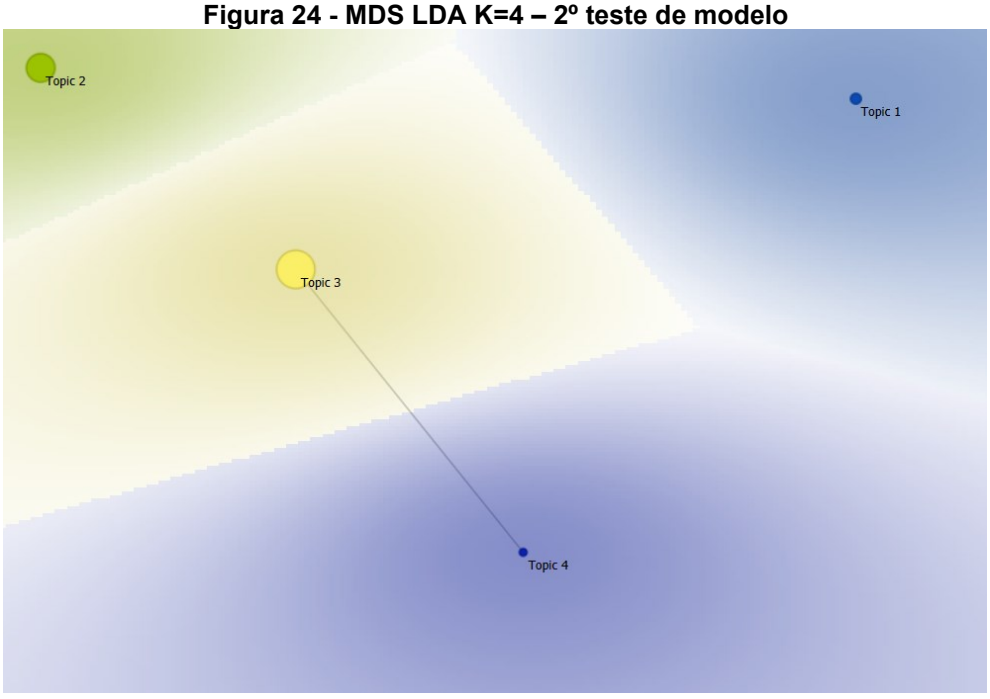
Reduzindo a quantidade de K tópicos para 5, e posteriormente 4, o panorama de proximidade/afinidade de tópicos foi conforme mostrado nas Figuras 30 e 31, seguintes:

Figura 23 - MDS LDA K=5 – 2º teste de modelo



Fonte: dados de pesquisa (2022)

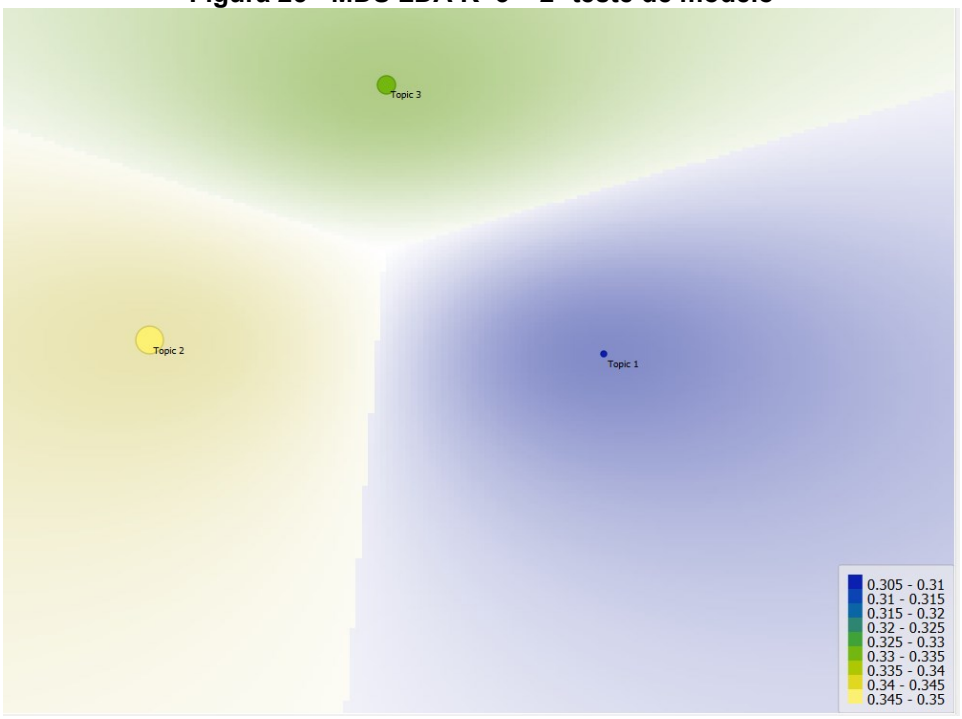
Figura 24 - MDS LDA K=4 – 2º teste de modelo



Fonte: dados de pesquisa (2022)

Foi verificado que, a similaridade indicada em K=5, pelos tópicos presentes na parte mais inferior da Figura 30, também foi demonstrada entre os Tópicos 3 e 4 da Figura 31, com K=4. Assim, de modo a poder formar tópicos mais distintos entre si, uma nova tentativa com K=3 foi conduzida, e o resultado é indicado na sequência, na Figura 32.

Figura 25 - MDS LDA K=3 – 2º teste de modelo



Fonte: dados de pesquisa (2022)

Verifica-se, pela visualização com a MDS, a predominância dos tópicos 2 e 3. A próxima etapa, são verificadas as *Keywords* dos tópicos e lidos os *tweets* mais representativos de cada tópico (**passo 11** e **passo 12**, respectivamente) e suas *Keywords*, sendo feita, então, a rotulação dos tópicos (**passo 13**), conforme abaixo:

- TÓPICO 1 – Cultura (virada cultural); Mobilidade pública.
- TÓPICO 2 – Custo de vida/Aluguel; Segurança pública.
- TÓPICO 3 – Mobilidade; Educação (piso professores); Sáude pública.

Ainda, para observação da correta divisão tópica, fez-se uso, neste segundo teste de aplicação do modelo, de um widget novo do Orange, denominado LDAvis[54], trazido do trabalho de Sievert e Shirley (2014). O LDAvis tem o propósito de permitir

[54] Acesso em 05 de Janeiro de 2023. Disponível em: https://orangedatamining.com/blog/2022/2022-03-18-ldavis/

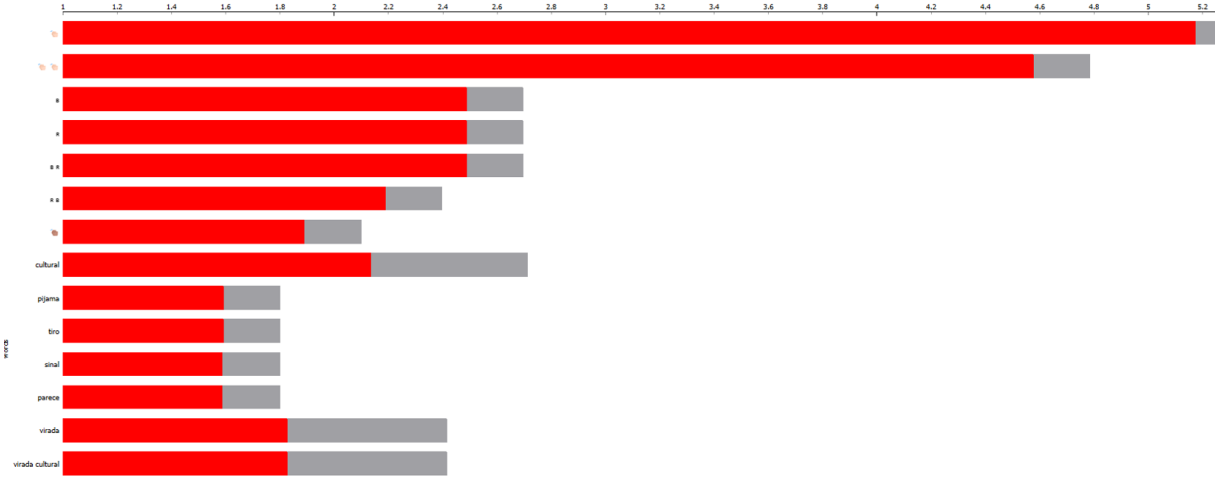
a visualização das relações entre tópico e termo, e, ao mesmo tempo, permitir uma visão geral dos tópicos, considerando prevalências e similaridades entre si (SIEVERT; SHIRLEY, 2014).

Assim, por meio deste, se faz possível a visualização das palavras mais significativas de cada tópico, apresentadas junto de seu coeficiente de relevância dentro do tópico, acrescidas, ainda, da informação de sua presença considerando o corpus de documentos como um todo.

A referida visualização é útil para que se possa verificar termos que de fato representem significativa e semanticamente cada tópico, e ainda verificar o quão são significativos para o dataset como um todo, não sendo exclusivamente relevantes a um tópico específico.

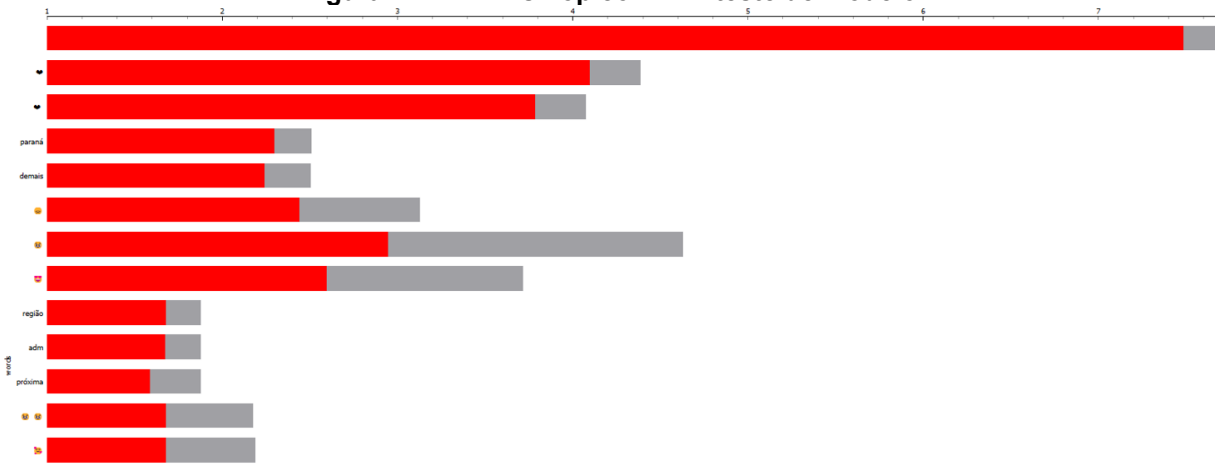
Para a identificação dos tópicos neste segundo teste de modelo, o panorâma mostrado no LDAvis, são as que constam nas figuras 33, 34 e 35, seguintes, sendo observadas em vermelho as frequências do termo dentro do tópico, e em cinza a frequência geral do termo, considerando o *corpus* de documentos alimentado.

Figura 26 - LDAvis Tópico 1 – 2º teste de modelo



Fonte: dados de pesquisa (2022)

Figura 27 - LDAvis Tópico 2 – 2º teste de modelo



Fonte: dados de pesquisa (2022)

Figura 28 - LDAvis Tópico 3 – 2º teste de modelo

Fonte: dados de pesquisa (2022)

Considerando a rotulação dos tópicos, verifica-se que o Tópico 1 possui termos associados à cultura, a exemplo de “cultural”, “virada” e “virada cultural”, além emojis que emulam palmas, sendo todos estes, consideravelmente exclusivos e delineados para este Tópico, do mesmo modo que o segundo tópico mais prevalente, qual seja, o Tópico 3, tem associadas palavras que remetem à mobilidade urbana, a exemplo de “contorno”, “contorno norte”, “norte, “zona”, além de outros termos que podem indicar associação à mobilidade. Diferentemente do Tópico 1, contudo, apesar de serem termos mais frequentes no Tópico 3, também estão presentes com razoável importância no *corpus* total de documentos, o que pode ser entendido como um tema prevalente no *dataset*.

Assim, tendo em vista ter sido identificado em dois dos três tópicos, além de ser tema de preocupação de gestores públicos brasileiros, no geral (CARVALHO, 2016), e tendo em vista, também, pouco antes do período do intervalo de coleta, ter havido a apresentação do Plano de Mobilidade Urbana de Maringá[55], é que se reputa adequada a escolha da temática da Mobilidade Urbana.

Então, na sequência, assim como feito no primeiro teste de modelo, fora formada uma lista de *Keywords* (**passo 14**) representativas do tema Mobilidade urbana, sendo utilizada como *input* para iterações na *Twitter* API, para extração temática de dados. As referidas *Keywords* são as que constam no Quadro 9.

Quadro 9 - Keywords temáticas Mobilidade

ônibus OR moto OR motoqueiro OR caminhão OR carro OR rua OR avenida OR multa OR transporte OR cruzar OR cruzamento OR vaga OR ciclo OR pista OR rodovia OR sinal OR semáforo OR acidente OR buraco OR mobilidade OR contorno OR viaduto OR passagem OR passageiro OR estacionar OR estacionamento OR estaciono OR estacionado OR aslfato OR aslfaltar OR asfaltado OR asfaltada OR calçada OR pedestre OR pedestres OR trafego OR trânsito OR via OR ciclovia OR ciclista OR bicicleta OR bike

Fonte: dados de pesquisa (2022)

Para iteração na *Twitter* API, quanto ao **passo 16** (de pesquisa geolocalizada), a referida lista de *Keywords* foi dividida em dois *scripts* em função da limitação de argumentos possíveis para extração de dados, tendo em mente se estar utilizando a API gratuita. Para o **passo 15**, de pesquisa utilizando-se das menções aos usuários centrais, não houve necessidade de separação em dois *scripts*.

Salienta-se que, uma vez mais, *tweets* advindos de contas de cunho meramente informativo, tais como as citadas no primeiro teste de modelo, foram desconsiderados.

Para contornar as já citadas limitações temporais da *Twitter* API, e conseguir resultados que considerassem um período maior do que 9 dias, os *scripts* dos passos 15 e 16 foram rodados semanalmente, geralmente, nas segundas-feiras, do dia 12/09 ao dia 19/12, compreendendo *tweets* produzidos entre o dia 04/09 e 19/12. Considerando um *overlapping* de datas, tendo em vista a extração ser semanal, a cada 7 dias, e serem extraídos documentos de até 9 dias anteriores, o montante total

[55] Acesso em 02 de Janeiro de 2023. Disponível em: http://www.maringa.pr.gov.br/site/noticias/2022/07/08/prefeitura-de-maringa-apresenta-plano-de-mobilidade-urbana-durante-conferencia-neste-sabado/40072

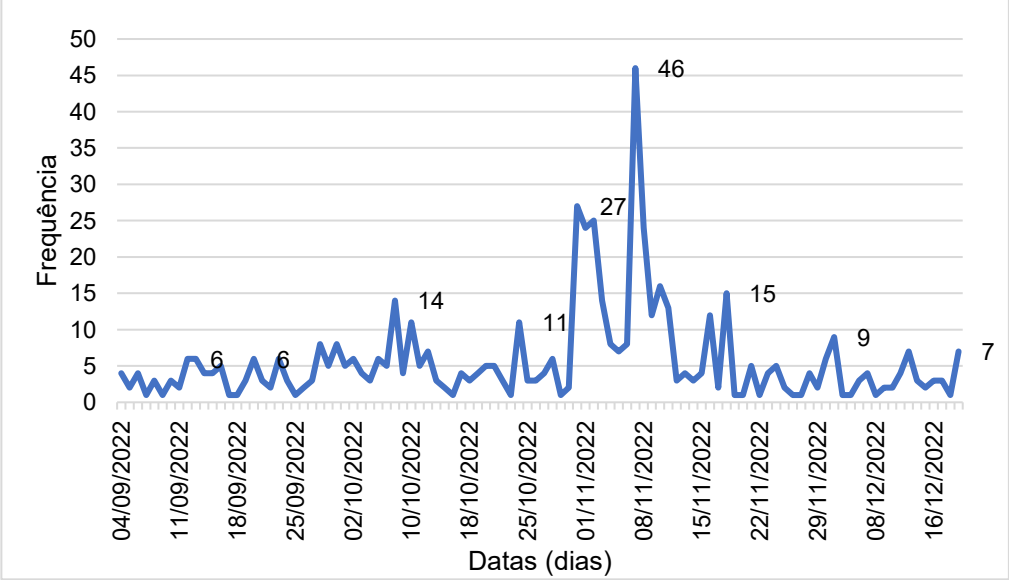
compilado de *tweets* inicialmente reunido resultou em 1587, para os quais, realizando a exclusão de duplicados, restaram – já desconsiderando *tweets* produzidos por contas meramente informativas – 916.

Destes 916, pela ampla variedade de *Keywords* utilizadas para extração, foi observado retorno de *tweets* não relacionados à mobilidade, assim, foram anotados os documentos correlacionados a temática da mobilidade urbana, restando, ao final, 581 *tweets.* Desde já se comenta, para as próximas execuções do modelo, a necessidade de uma escolha mais delimitada/criteriosa de *keywords*, de modo a evitar a coleta ruidosa de dados. Vale lembrar que, trabalhos com coleta temática primária, contaram com a ajuda de especialistas nos assuntos para definição das palavras-chaves mais adequadas, a exemplo de Musto et al. (2015), que contaram com a ajuda de psicólogos experientes para definição de *keywords* relacionadas à discursos de ódio.

Considerando o *corpus* de 581 documentos, na sequência fez-se a importação ao *Orange* (**passo 17**) para posterior pré-processamento (**passo 18**).

O Gráfico 4 mostra a distribuição dos documentos extraídos, considerando o período compreendido no *dataset*.

Gráfico 3 - Frequência diária de tweets no Dataset – 2º teste de modelo



Fonte: dados de pesquisa (2022)

A média de postagem diária, no período, foi de 5,7 *tweets*, tendo a amostra um devio padrão de 6,59. A média foi menor do que a identificado no primeiro teste de modelo, muito em virtude de não ter havido, em regra, a situação excepcional de

calamidade como a havida no caso da saúde pública em janeiro de 2022, fruto da pandemia do Covid. Ainda, a variabilidade de postagens diárias também foi menor do que no primeiro teste de modelo, indicando maior uniformidade e continuidade de interesse dos *netizens* de Maringá quanto à temática

As exceções patentes à uniformidade comentada, podem ser verificadas nos dias 31/10 e 07/11, sendo este último, o dia em que houve o pico de frequência no conjunto de dados. Pela análise dos *tweets* acredita-se o majorado volume de postagens ser devido, respectivamente, às movimentações – sobremaneira em trechos de rodovia – frente aos resultados do segundo turno das eleições presidenciais, e manifestações de contestação aos resultados das eleições, obstruindo o fluxo, sobretudo na Avenida Mandacaru. Notícias sobre estes ocorridos podem ser verificadas aqui[56] [57].

Desde já se verifica a possibilidade de atuação tempestiva, pelo poder público, tendo em mãos informações pontuais de acontecimentos em locais indicados pelos *netizens*, como exemplo das obstruções na Avenida Mandacarú. Uma possibilidade para estudos futuros poderia ser a tentativa de implementação/adaptação do modelo para contemplar *feedbacks* em tempo real, de maneira mais automatizada.

Ainda, não obstante não constar no fluxo do modelo, de modo à agregar à análise descritiva do sentimento (**passo 23**) fora aplicada a LDA no *dataset* temático de Mobilidade, de modo a poder ter uma análise mais automatizada, de *machine learning*, acerca dos microtemas identificados dentro da temática da Mobilidade urbana, dentro do período compreendido. Os tópicos identificados foram:

- TÓPICO 1 - PARALISAÇÕES/MANIFESTAÇÕES ; MANUTENÇÃO VIÁRIA
- TÓPICO 2 - ATOS RUA MANDACARU ; MANUTENÇÃO VIÁRIA
- TÓPICO 3 - MANUTENÇÃO VIÁRIA ; RELATOS DE INFRAÇÕES DE TRÂNSITO

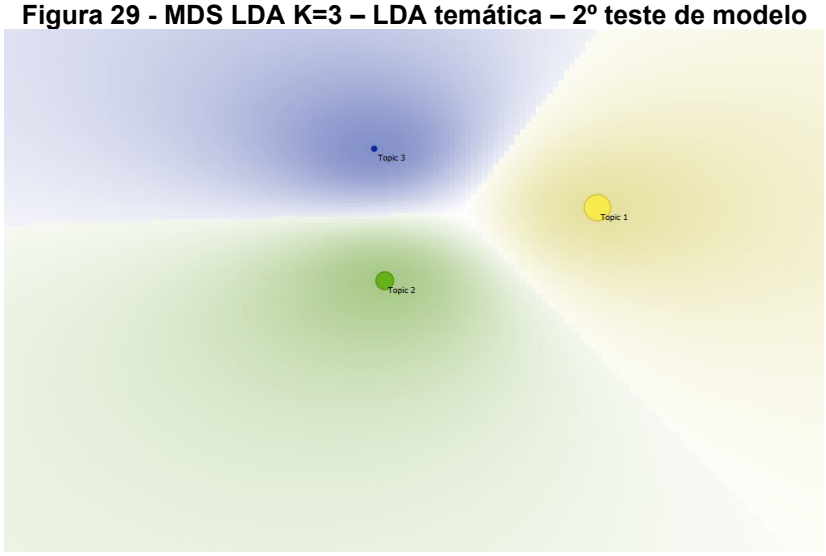
A distribuição tópica final, representada graficamente no MDS consta conforme Figura 36.

---

[56] Acesso em 04 de Janeiro de 2023. Disponível em: https://g1.globo.com/pr/norte-noroeste/noticia/2022/10/31/manifestantes-bolsonaristas-bloqueiam-rodovia-em-protesto-a-vitoria-de-lula-em-maringa.ghtml

[57] Acesso em 04 de Janeiro de 2023. Disponível em: https://www.cbnmaringa.com.br/noticia/transito-na-avenida-mandacaru-esta-fluindo-normalmente-diz-secretario

Figura 29 - MDS LDA K=3 – LDA temática – 2º teste de modelo



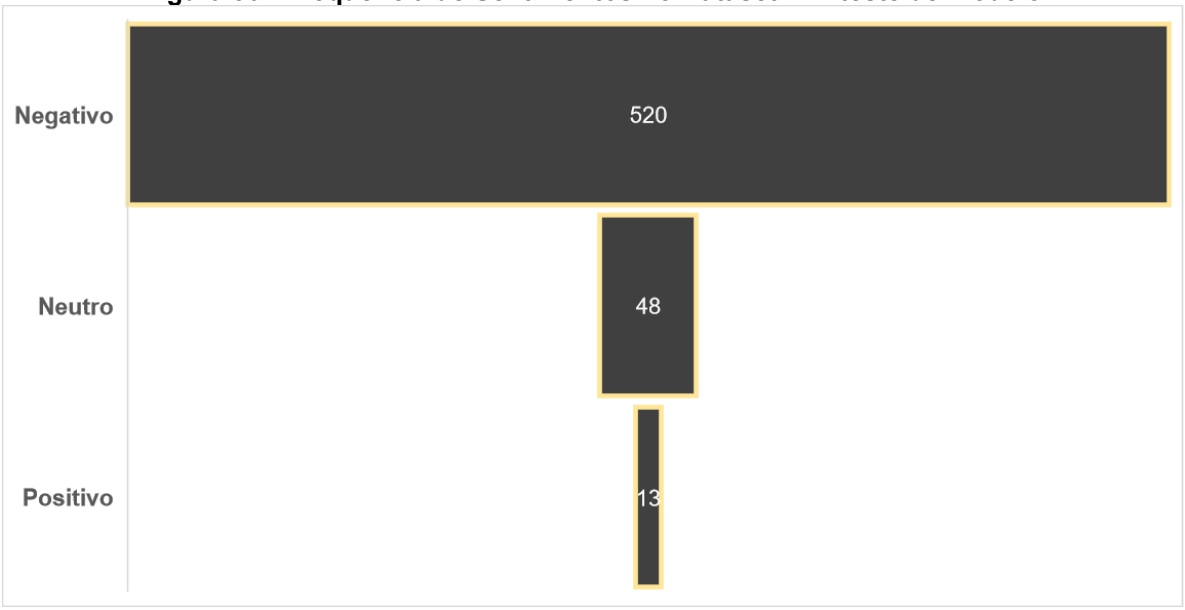
Fonte: dados de pesquisa (2022)

Foram seguidas as mesmas lógicas de visualização e rotulagem conforme já explicado para ambos os testes de modelo, razão pela qual não serão explicitamente repisadas aqui. Foi verificado que, ao encontro da análise do conteúdo e motivação das frequências de postagens, feita quando analisados os picos diários de produção de *tweets*, o Tópico 2 também trouxe os ocorridos na avenida Mandacarú, bem como fora observado a pujança do Tópico 1 quanto às manifestações, também verificada na análise manual dos picos de frequência.

Interessante notar que, não identificado nos picos de frequência, aparece o assunto da manutenção viária, frequente no *dataset* temático, e que, apesar de constante, não causou *burst* de *tweets*, o que pela análise “manual”, não fora identificado. A identificação pela LDA temática mostra que a inclusão deste passo deve ser considerada no fluxo do modelo, não só para identificação de subtópicos, mas também por se tratar de um modo automatizado, que proporciona celeridade na identificação dos assuntos intratópico.

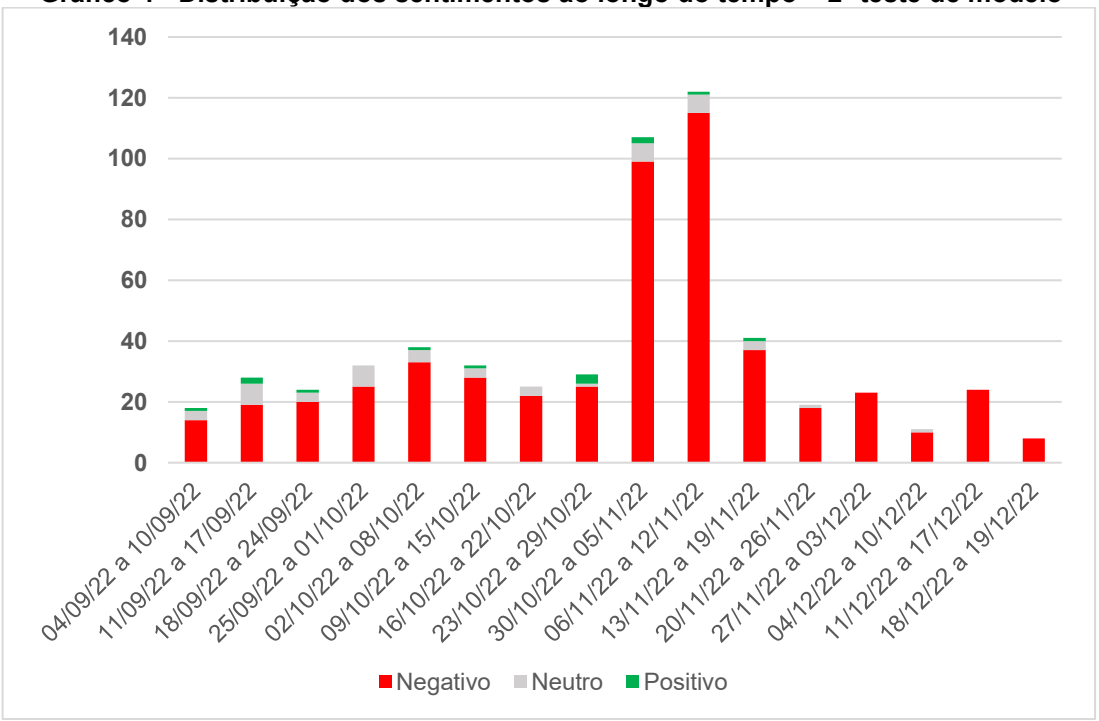
Tendo havido a marcação dos *tweets* para análise de sentimentos, **passo 19**, o panorâma obtido foi o que consta nos Gráficos 5 e 6.

Figura 30 - Frequência de Sentimentos no Dataset – 2º teste de modelo



Fonte: dados de pesquisa (2022)

Gráfico 4 - Distribuição dos sentimentos ao longo do tempo – 2º teste de modelo



Fonte: dados de pesquisa (2022)

Tendo em vista o período compreendido, o Gráfico 6 (**passo 25**) está sendo apresentado, para mais limpa visualização, semanalmente.

Uma vez mais, tal como no primeiro teste de modelo, o sentimento mais frequentemente observado, e desta vez de maneira mais acentuada, foi o sentimento negativo, o que, uma vez mais, reforça a idea de que geralmente os netizens utilizam o twitter para compartilhar suas frustrações, reclamações e afins (EL-DIRABY; SHALABY; HOSSEINI, 2019).

Referente ao desempenho do SVM e do RF, para a classificação do sentimento, primeriamente, faz-se, novamente, a explicação de que, como foram os que tiveram os melhores desempenhos no primeiro teste de modelo, foram os mantidos para a rodagem do segundo teste.

Tal como feito no primeiro teste de modelo, e como recomendado por Kazmaier e Vuuren (2020), diferentes parâmetros para verificação do desempenho foram testados, qual seja, neste caso, rodar com e sem pré-processamento, de modo a verificar a diferença de *performance,* nos resultados. As demais escolhas para operacionalização, como já explicadas no primeiro teste de modelo, não serão repetidas aqui.

Constam na sequência (representando o **passo 30**), os resultados com e sem pré-processamento, respectivamente:

Com pré-processamento, conforme Tabela 10:

Tabela 10 - Desempenho com pré-processamento – 2º teste de modelo

Classificador	AUC	CA	F1	Precision	Recall
SVM	1.000	0.895	0.845	0.801	0.895
Random Forest	1.000	0.981	0.980	0.981	0.981

Fonte: dados de pesquisa (2022)

Sem pré-processamento, conforme Tabela 11:

Tabela 11 - Desempenho sem pré-processamento – 2º teste de modelo

Classificador	AUC	CA	F1	Precision	Recall
Random Forest	0.655	0.922	0.892	0.911	0.922
SVM	0.694	0.914	0.873	0.835	0.914

Fonte: dados de pesquisa (2022)

Desempenhos mais consistentes em todos os indicadores, considerando, desta vez, o pré-processamento. Mais especificamente, se for tomado o indicador AUC para análise, é verificado que os classificadores, com o pré-processamento, foram exitosos em determinar a melhor “curva” ou *threshold* que classifica o *dataset*,

o que pode indicar que, com o pré-processamento, o conjunto de dados tornou-se mais homogêneo, favorencendo a classificação. Mais comentários serão apresentados posteriormente.

Com relação aos desempenhos por categoria de sentimento, tem-se, conforme abaixo.

- Com pré-processamento:

a) Desempenho para classificação de *tweets* com sentimento Negativo, conforme Tabela 12:

Tabela 12 - Desempenho com pré-processamento para classificação de tweets negativos – 2º teste de modelo

Classificador	AUC	CA	F1	Precision	Recall
Random Forest	0.883	0.922	0.959	0.922	1.000
SVM	0.737	0.914	0.955	0.914	1.000

Fonte: dados de pesquisa (2022)

b) Desempenho para classificação de *tweets* com sentimento Neutro, conforme Tabela 13:

Tabela 13 - Desempenho com pré-processamento para classificação de tweets neutros – 2º teste de modelo

Classificador	AUC	CA	F1	Precision	Recall
Random Forest	0.906	0.922	0.222	1.000	0.125
SVM	0.758	0.914	0.000	0.000	0.000

Fonte: dados de pesquisa (2022)

c) Desempenho para classificação de *tweets* com sentimento Positivo, conforme Tabela 14:

Tabela 14 - Desempenho com pré-processamento para classificação de tweets positivos – 2º teste de modelo

Classificador	AUC	CA	F1	Precision	Recall
Random Forest	0.482	0.922	0.000	0.000	0.000
SVM	0.768	0.914	0.000	0.000	0.000

Fonte: dados de pesquisa (2022)

- Sem pré-processamento:

d) Desempenho para classificação de *tweets* com sentimento Negativo, conforme Tabela 15:

Tabela 15 - Desempenho sem pré-processamento para classificação de tweets negativos – 2º teste de modelo

Classificador	AUC	CA	F1	Precision	Recall
Random Forest	0.638	0.922	0.959	0.922	1.000
SVM	0.693	0.914	0.955	0.914	1.000

Fonte: dados de pesquisa (2022)

e) Desempenho para classificação de *tweets* com sentimento Neutro, conforme Tabela 16:

Tabela 16 - Desempenho sem pré-processamento para classificação de tweets neutros – 2º teste de modelo

Classificador	AUC	CA	F1	Precision	Recall
Random Forest	0.659	0.922	0.222	1.000	0.125
SVM	0.675	0.914	0.000	0.000	0.000

Fonte: dados de pesquisa (2022)

f) Desempenho para classificação de *tweets* com sentimento Positivo, conforme Tabela 17:

Tabela 17 - Desempenho sem pré-processamento para classificação de tweets positivos – 2º teste de modelo

Classificador	AUC	CA	F1	Precision	Recall
Random Forest	0.717	0.922	0.000	0.000	0.000
SVM	0.772	0.914	0.000	0.000	0.000

Fonte: dados de pesquisa (2022)

São observados, no geral, desempenhos melhores do que no primeiro teste de modelo. Este fator pode ser devido à alguns motivos. Um deles perpassa a perícia do autor deste trabalho, que, não sendo expert em categorizar/classificar sentimento, mas tendo, ao longo do período de desenvolvimento deste trabalho, realizado anotação manual de sentimento em *datasets* distintos, continuamento, acredita-se que tenha havido uma evolução na assertividade de rotulação dos sentimentos dos *tweets*. Ainda, e ainda mais significativo, é o fato de que a temática da mobilidade urbana, sendo um problema importante nas cidades (CARVALHO, 2016) suscita a explicitação de comentários em tom mais crítico e descontentamentos do que elogios,

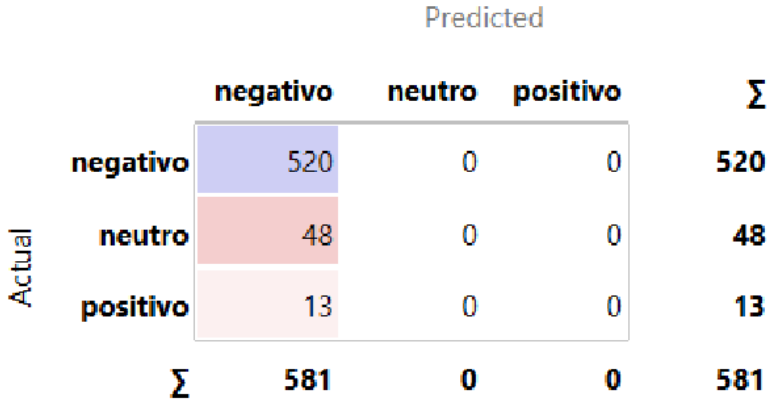
por exemplo, o que é traduzido pela quantidade majoritária de *tweets* classificados como “negativo”, assim, os resultados de classificação podem ter sido favorecidos em razão de haver esta disparidade na proporção de categorias de sentimento, dentro do conjunto de dados.

Para melhor visualização e para cumprimento da apresentação gráfica do desempenho, são trazidas, abaixo, as *Confusion Matrixes* (representando o **passo 29**).

- Com pré-processamento:

c) SVM, conforme Figura 45:

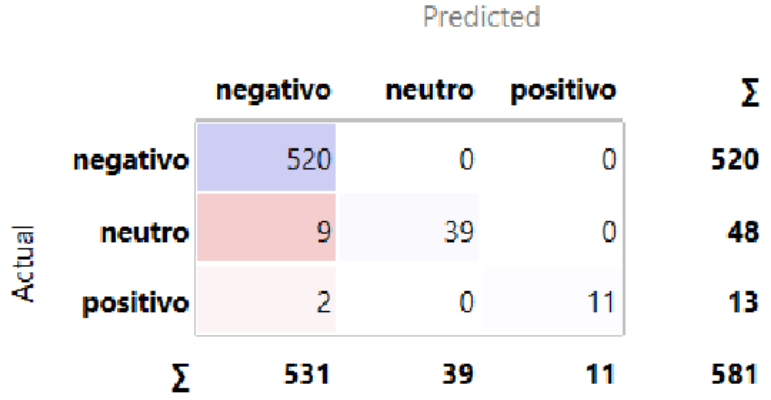
Figura 31 - Desempenho SVM com pré-processamento – 2º teste de modelo



Fonte: dados de pesquisa (2022)

d) *Random Forest,* conforme Figura 46*:*

Figura 32 - Desempenho Random Forest com pré-processamento – 2º teste de modelo

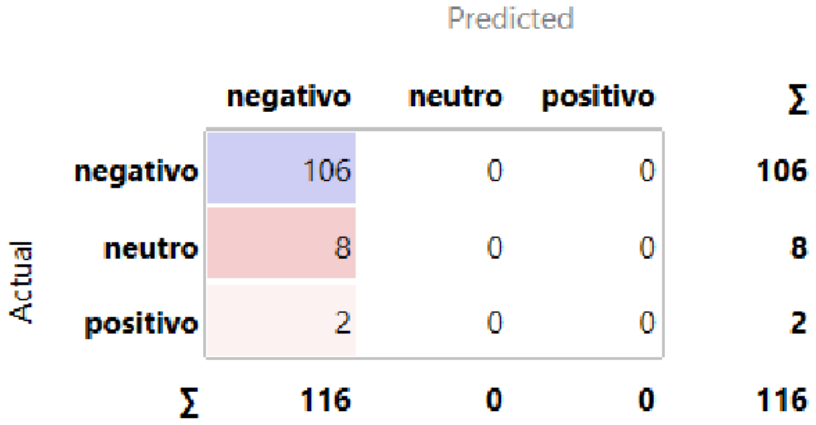


Fonte: dados de pesquisa (2022)

- Sem pré-processamento:

c) SVM, conforme Figura 47:

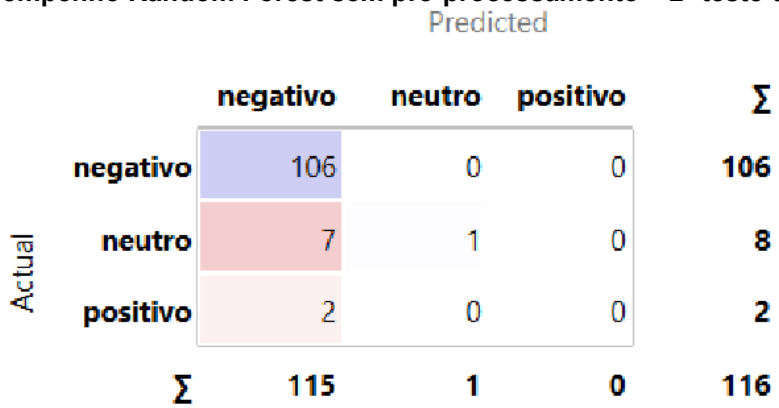
Figura 33 - Desempenho SVM sem pré-processamento – 2º teste de modelo



Fonte: dados de pesquisa (2022)

d) Random Forest, conforme Figura 48*:*

Figura 34 - Desempenho Random Forest sem pré-processamento – 2º teste de modelo



Fonte: dados de pesquisa (2022)

Neste segundo teste de modelo o classificador com melhor desempenho foi o *Radom Forest*, seguido do SVM, o que pode ser atestado, principalmente quanto à capacidade de classificação de *tweets* neutros e positivos, bem-sucedida no primeiro e ao revés no segundo, conforme figuras 45 e 46.

Passao ao **passo 26**, foram feitas as *wordclouds* seguintes, para cada categoria de sentimento, tendo havido uma etapa de pré-processamento, para limpeza dos dados.

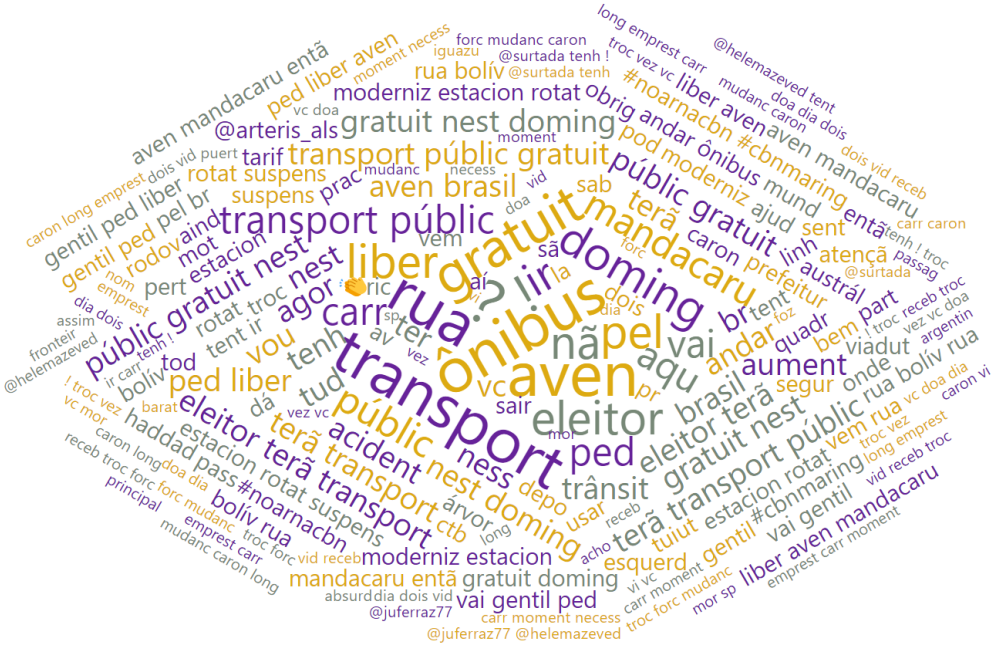
- *Word-Cloud* dos *tweets* classificados como Negativo, conforme Figura 49:

Figura 35 - Word-Cloud dos tweets classificados como negativo – 2º teste de modelo

Fonte: dados de pesquisa (2022)

- *Word-Cloud* dos *tweets* classificados como Neutro, conforme Figura 50:

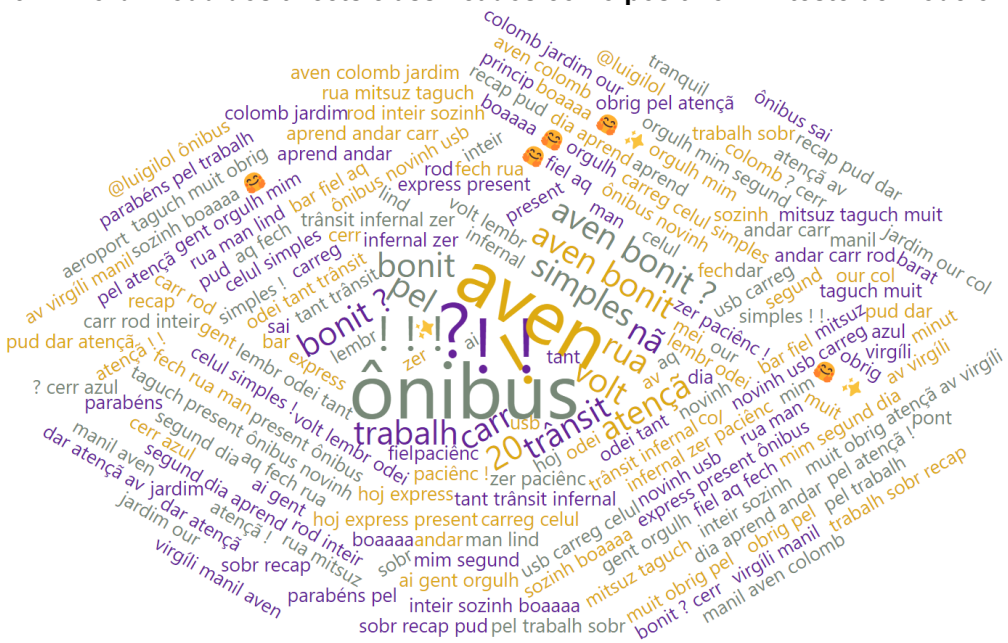
Figura 36 - Word-Cloud dos tweets classificados como neutro – 2º teste de modelo



Fonte: dados de pesquisa (2022)

- *Word-Cloud* dos *tweets* classificados como Positivo, conforme Figura 51:

Figura 37 - Word-Cloud dos tweets classificados como positivo – 2º teste de modelo



Fonte: dados de pesquisa (2022)

Primeiramente se observa, haja vista a extração temática, haver palavras que se repetem nas três nuvens, entretanto, indicativo do sentimento, por exemplo, é a inexistência de emojis, para a nuvem de sentimento negativo, ao passo que constam na nuvem de sentimento positivo, além de outros termos como “bonit” (reduzido de bonito, fruto do processo de *stemming*), e de termos negativos – na *word cloud* negativa – como “vergonha”, “palhac” e “bardern”, também palavras “stemizadas”. A *wordcloud* de sentimento neutro é menos acentuada, e não apresenta palavras com muito significado positivo ou negativo, por si só, decorrente de ser fruto de comentários amenos, e, por vezes, dúvidas legítimas direcionadas aos usuários centrais, como se o transporte público seria gratuito, no dia de votação, pelo que pode se depreender.

Justifica-se, novamente, na sequência, a não realização de partes das análises inicialmente propostas e realizadas no primeiro teste de modelo.

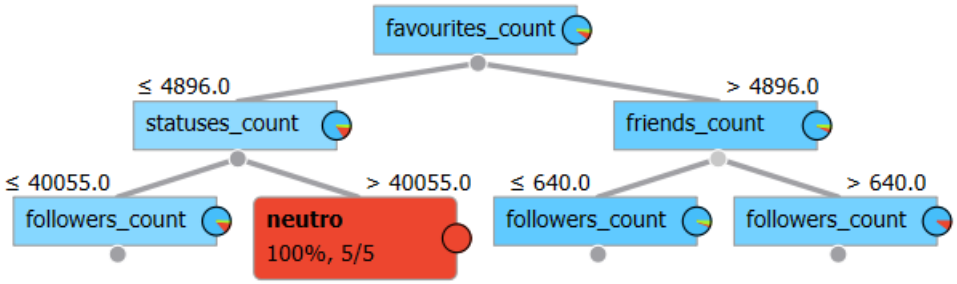
Uma vez mais menciona-se a não realização do passo 20, de análise não supervisionada do sentimento, por não ter sido frutífera, no primeiro teste de modelo.

Referente à análise de emoções (passo 21), de maneira não supervisionada, também se mostrou infrutífera. A estratégia supervisionada poderia ser adotada, com a anotação manual das emoções de cada *tweet*, contudo, tal tarefa, para ser efetiva,

deve ser realizada prioritariamente por especialista na temática que se deseja atacar, a exemplo do que fizeram Musto et al., (2015), em que se utilizaram psicólogos experientes para isso. Outra saída, seria encontrar na literatura revisada, um dataset classificado em português, que já tenha sido testado, o que não fora encontrado, ficando sua construção, desde já, como possibilidade para estudos futuros.

Referente à visualização da relação dos metadados frente ao sentimento (**passo 24**), temos, conforme abaixo, Figura 52:

Figura 38 - Diagrama de árvore relações entre metadados dos tweets – 2º teste de modelo



Fonte: dados de pesquisa (2022)

Novamente, a metadado mais influente sobre a polaridade de sentimento no conjunto de dados foi o “favourites_count”, cuja definição já foi trazida na seção anterior, do primeiro teste de modelo.

Contudo, a análise aqui é a de que a árvore de decisão pode não ter uma usabilidade muito grande pelos gestores públicos, na medida em que não implica em dados nos quais se possa agir, objetivamente, mas requer interpretação que serve mais para um entendimento aprofundado de fatores de certo modo “reprimidos” ou “escondidos”, que podem influenciar no sentimento.

Ainda, quando se tem *datasets* com sentimentos predominantes, como é o caso do prsente trabalho, a análise acaba se tornando pouco agregadora, na medida em que os metadados podem acabar não refletindo de forma correta, sua importância para a determinação do sentimento, funcionado, ao revés, de maneira mais adequada, caso o conjunto de dados fosse equilibrado na proporção do sentimento, ficando, assim, como oportunidade para que isso seja feito, isoladamente, ao menos, em trabalhos futuros. Assim, para a construção do fluxo final do modelo, é uma análise a ser desconsiderada.

4.2.4 Implicações Práticas para a Gestão Pública

Esta seção tem o intuito de enfatizar, de maneira clara, os aspectos práticos para a gestão das cidades, que a aplicação do modelo vem suscitar.

Embora não previsto originalmente, traz-se aqui aquele que seria, considerando o fluxo original, o **passo 32,** cujo propósito é o de unir a apresentação de todos os resultados, indicando a necessidade do gestor/operador, que fará uso das informações apresentadas, de interpretar os dados e propor ações à administração, no sentido de corrigir o que não está a contento, por exemplo. Segue-se lógica parecida, quanto às recomendações, ao que fora apresentado no trabalho de Adamu et al., (2021).

São apresentadas possibilidades de utilização dos dados e análises aqui apresentados para a gestão pelo poder público.

Primeiramente, a utilização se dá no domínio de entender as principais preocupações dos *netizens* da cidade que está fazendo uso do modelo, externalizadas, neste caso, pelo(s) tópico(s) principal(is) identificado(s) pela LDA. Esta identificação fornece mais uma fonte de informação como subsídio ao poder público poder priorizar áreas de investimento, por exemplo.

A tempestividade do modelo, o qual poderia, inclusive, servir de base para a implementação de coletas e análises em tempo real – a depender do nível de acesso conseguido, junto a API – é capaz de indicar disrupções no ambiente das cidades (similar ao que fizeram Fan, Jiang e Mostafavi, (2021), tais como os presentes no *dataset*, a exemplo das reclamações acerca das manifestações e obstruções de fluxo na Avenida Mandacarú. *Tweets* Exemplificativos:

@UlissesMaia por favor prefeito! Fecharam a avenida Mandacaru e minha rua! Algo tem que ser feito!
lembrando aqui q quando foi cancelado carnaval e um monte de jovem foi pro Willie Davis e um monte de polícia foi lá tirar a gente e AGR até a semob foi fecha o trânsito na mandacaru, EAE @UlissesMaia vai resolver essa papagaiada não

Ademais, indicações de ruas que precisem de manutenção viária também podem ser abstraídas dos dados obtidos, sendo um dos assuntos presentes na modelagem de tópicos temática feita neste segundo teste de modelo.

@UlissesMaia @CaiqeRams Início da rua Bolívia indo pela colombo até ali na São Domingos é IMPOSSÍVEL, um buraco maior que o outro em toda extensão, não sei a continuidade dela mas ta feia a situação

@UlissesMaia E os buracos na rua Alfinete inserido https://t.co/oyFewBpsZV Vão tapar quando? Tá dando pra planta uma bananeira já

Isto posto, aliado a outros instrumentos da tecnologia da informação e comunicação como o *Google Street View*[58], e tem-se uma abordagem de baixo custo para identificação de ruas que precisem de recapamento, por exemplo, sem haver a necessidade de visita em lócus para tanto.

Ainda mais útil teria sido a utilização destes dados quando da época de maior crise da Pandemia da Covid-19, quando foi rodado o primeiro teste de modelo. A identificação das reações dos *netizens* perante a atuação dos gestores públicos, e a pandemia de um modo geral, é um impulso para mudança de rota ou atendimento a demandas específicas, durante o período pandêmico. Ainda, com a possibilidade da LDA temática, a identificação de subtópicos pode indicar áreas que precisam de mais atenção do que outras, a exemplo de fila nos atendimentos ou falta de testes rápidos.

Uma vez mais, independentemente da temática, considera-se uma possibilidade futura a utilização deste modelo como fundamento à operacionalização em tempo real, para maior tempestividade da atuação da administração pública.

Assim, tal qual feito em Adamu et al., (2021), e, tomando por base o Segundo Teste de modelo, sugere-se, em posse dos dados analisados, as seguintes ações:

1. Que seja instituído, a partir do modelo e estratégias de extração de dados propostos, um sistema de monitoramento de *tweets*, onde sejam identificadas disrupções – tal como Fan, Jiang e Mostafavi (2020) – de mobilidade, que citem localidades específicas, de modo a poder haver a atuação tempestiva do poder público frente às situações causadoras das disrupções;
2. Que sejam instalados (ou fortalecidos, se já existentes) programas de manutenção viária, que considerem os dados extraídos e apresentados pelo modelo proposto, sobretudo tendo em vista as vias já indicadas e seus problemas respectivos, de modo a promover melhorias na mobilidade urbana e amenizar a insatisfação dos *netizens* perante este assunto.

Ainda, sob a lógica das cidades inteligentes, levar em conta os anseios dos cidadãos, por meio de suas opiniões exaradas em redes sociais intersecta não só a

[58] Acesso em 09 de Janeiro de 2023. Disponível em: https://www.google.com/intl/pt-BR/streetview/

dimensão de governo inteligente, no que diz respeito à sua participação, como também a própria dimensão de *smart people* (GIFFINGER et al., 2007), ou população inteligente – em tradução livre – quando se entende que levar em conta suas opiniões eleva a própria interação social que eles tem consigo mesmo, bem como junto das autoridades públicas, que são, no caso deste trabalho, os próprios usuários centrais aos quais são dirigidas as mensagens.

Reforça-se também, o pressuposto de atendimento das demandas dos usuários dos serviços das cidades, por meio do uso da TIC (BATTY et al., 2012), sobremaneira tendo em vista o fato de dados de redes sociais serem um patente subgrupo das ferramentas de TIC.

A aplicação do modelo permite uma fonte que pode ser, por si só, útil à gestão pública, na busca dos anseios mais emergenciais da sua população (a exemplo da saúde pública e mobilidade urbana, resultados das duas aplicações do teste de modelo), mas que tem ainda maior potencial, se utilizada em conjunto com outras fontes de dados. Tal fator foi observado e recomendado na literatura revisada, não só pra uma tomada de decisão mais bem informada, aumentando a consciência situacional (FAN; JIANG; MOSTAFAVI, 2020), como também para reduzir o enviesamento da utilização de uma só fonte de dados (ABDUL-RAHMAN et al., 2020).

Nesse sentido recomenda-se que o modelo seja tomado em conjunto com outras fontes de dados, a exemplo do sistema de monitoramento utilizando câmeras, tal como já existente na cidade de Maringá, lócus da aplicação dos testes de modelo[59]. A utilização do sistema de monitoramento por câmeras pode confirmar indicações textuais dos cidadãos, extraídos como resultado do modelo, ou mesmo fazer a utilização dos dados de redes sociais como insumos para direcionar instalações de câmeras em regiões/ruas que necessitam mais, conforme citadas nos *tweets* coletados, similar ao que fizeram Costa et al., (2018), onde os autores propuseram uma maneira de otimizar o monitoramento por meio da atribuição de prioridades de monitoramento aos sensores, com base em dados do *Twitter.* Uma priorização possível, trazendo aos resultados dos testes de modelo, seria a instalação/direcionamento de câmeras para a avenida mandacaru, bastante citada

[59] Acesso em 08 de março de 2023. Disponível em: http://www.maringa.pr.gov.br/site/noticias/2022/09/02/prefeitura-inicia-instalacao-de-70-cameras-de-monitoramento-para-identificacao-facial-e-leitura-de-placas-de-veiculos/40348

conforme segundo teste de modelo, ou avenida JK, onde a existência de um buraco na via foi relatada.

A indicação de possíveis barreiras que dificultem a mobilidade urbana fora relatada em outro trabalho da literatura (SÁNCHEZ-ÁVILA et al., 2020), cujo lócus de aplicação foi a Espanha. Trazendo ao contexto dos presentes resultados das aplicações do modelo, sugere-se o melhoramento da utilização da marcação de geolocalização dos *tweets* e detecção das menções de ruas/avenidas no texto dos *tweets*, para melhor usabilidade no contexto das cidades, além de adaptar o modelo para tempo real, de modo a revestí-lo de maior tempestividade.

Comenta-se, não só sobre as implicações práticas e possíveis ações a serem tomadas, de posse das informações resultantes do modelo, mas também quanto ao papel dos gestores públicos, como o prefeito, usuário central por meio do qual fora feita a coleta, e seu papel na influência do sentimento público. Em Alkhatib et al., (2020) verificou-se que os líderes de opinião, em seu trabalho, tinham o poder de exercer grande influência na opinião pública, com alguns *tweets,* sobremaneira verificada a redução de opositores. Isto deve ser levado em mente e considerado dentro de uma estratégia de comunicação do prefeito, com sua população. *Tweets* com o objetivo de acalmar os ânimos, e prestar conta das ações em curso para resolução de problemas, principalmente em situações de crise, como as relatadas nos testes de modelo, (frente ao combate da pandemia, por exemplo), pode contribuir na amenização do descontentamento/preocupação de seus cidadãos.

Por fim, na esteira da recomendação de Musto et al., (2015) para que, sendo o estudo conduzido de maneira mais longitudinal, sejam feitas as comparações temporais da evolução dos indicadores de sentimento, de modo a ter um panorâma da evolução dos cenários encontrados, e acompanhamento das ações de correção tomadas.

Tendo tratado dos resultados encontrados, em ambos os testes de modelo, passa-se, na sequência, ao capítulo de fechamento do presente trabalho, sem antes ser apresentado o modelo final, que tem por base o modelo incial, somado aos resultados de testes de modelo.

4.3 Apresentação do Modelo Final

Primeiramente, far-se-á a intitulação do modelo, que se chamará *City Data Scraper through Central Users* – CDSTCU, que, em tradução livre, significa “raspador” de dados das cidades por meio de usuários centrais, que, por razões estéticas e de amplitude/abrangência, será mantido em inglês.

Tendo em vista os resultados do primeiro teste de modelo, e considerando a usabilidade, pelo poder público, das informações obtidas com o que aqui é proposto, foi verificado que algumas das estratégias propostas acabavam sendo pouco frutíferas, razão pela qual já não foram executadas no segundo teste de modelo, ou, tendo sido executadas, foi verificada não serem tão adequadas ao propósito focal do trabalho, quais sejam, sobremaneira, a identificação dos tópicos de discussão pelos *netizens*, e a análise de sentimento destes. Assim, com apoio nas duas *frameworks* utilizadas para construção (CUP e ECCO), juntamente com a literatura revisada, e tendo sido analisados os resultados das aplicações dos dois testes de aplicação do modelo, é que é proposta a edição final do modelo, ou, modelo final CDSTCU.

Inicialmente, optou-se por manter na execução do CDSTCU somente 2 classificadores que melhor desempenharam no primeiro teste de modelo, razão pela qual os classificadores indicados pelos passos 19.1.1 e 19.1.2 foram desconsiderados.

O passo 20, inicial, tratava da análise de sentimento não supervisionada VADER, e, desde a primeira aplicação, já se supunha que não teria resultados adequados, visto ser tecida à conjuntos de dados de língua inglesa. Mesma lógica para o passo 21, que tratava da análise de emoções, a qual, não havendo abordagem não supervisionada, no *Orange*, para dados em português, necessitava de rotulação do *dataset*. Contudo, tal trabalho de rotulação não é simples, havendo trabalhos, como o de Musto et al., (2015), que, para definir *Keywords* de dicurso de ódio, contou com psicólogos experientes para realizar a tarefa. Assim, necessitando para categorização de conhecimentos específicos na identificação dos tipos de emoção que poderiam estar associados aos *tweets*, e não tendo encontrado *datasets* já testados, em português, a referida etapa não foi realizada, ficando como possibilidade para estudos futuros. Os passos 27 e 28 eram esquemas de visualização associados ao passo 20 e 21, razão pela qual foram desconsiderados.

O passo 24 foi realizado, contudo, ao final da própria seção anterior, já se argumenta quanto à utilidade na usabilidade pelo poder público do referido passo,

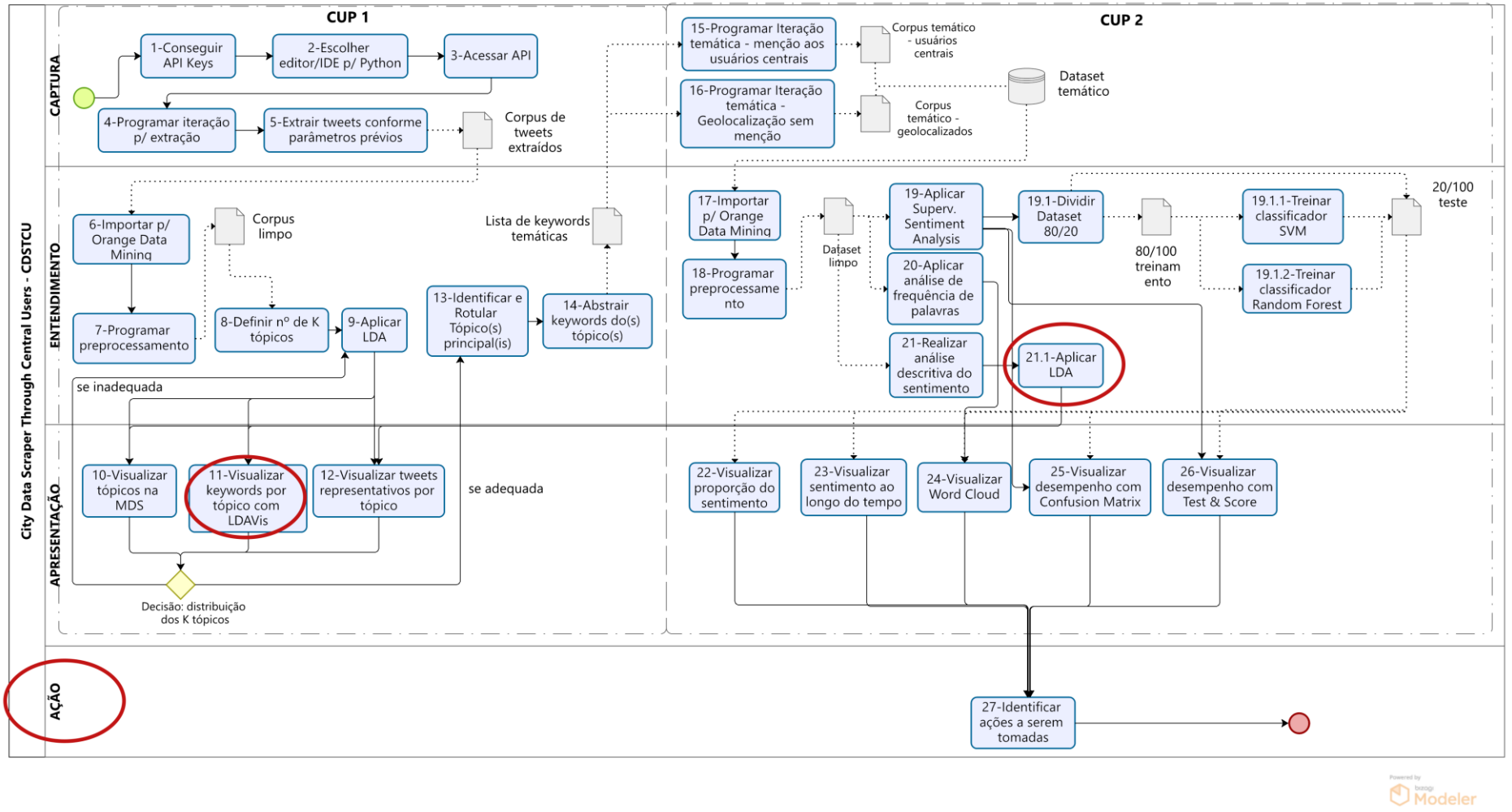
sendo útil, no entanto, que se tente estudar, isoladamente, as *features/*metadados que influenciam no sentimento de um *tweet*, para além do conteúdo textual.

O modelo atualizado contempla, a visualização das *Keywords* por tópico por meio do LDAvis, conforme aplicado no segundo teste de modelo, incrementada no passo 11. Agrega, também, uma análise realizada no segundo teste de modelo, qual seja, a aplicação da modelagem de tópicos/LDA, no conjunto de dados temático, com o fito de complementar/aprofundar a análise descritiva do sentimento – constando como um subpasso do passo 20 – e dos picos de frequência de postagem.

Finalmente, buscando uma maior internalização do aspecto de uso prático, inseriu-se uma última etapa de ações recomendadas, com base nos dados acessados, de modo a deixar mais explícita a utilidade para a gestão pública.

Fluxogramaticamente, o CDSTCU, de maneira completa e atualizada na Figura 53. É feita a marcação com elipses vermelhas de etapas/passos incluídos frente ao modelo inicial.

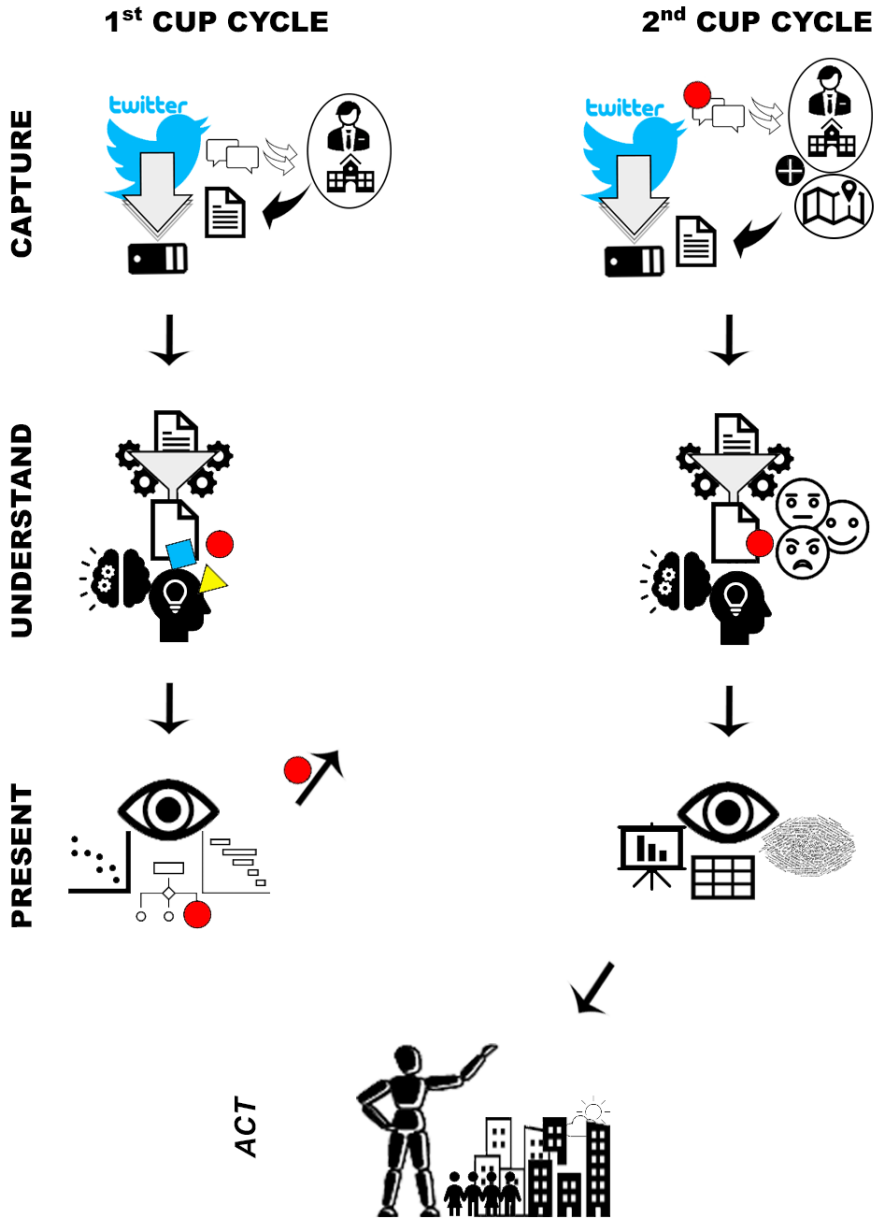
Figura 39 – CDSTCU - Framework proposta – Fluxo Atualizado



Fonte: dados de pesquisa (2022)

Entretanto, para que as propostas não realizadas/concluídas aqui, deixem de ser tentadas em estudos futuros, propõe-se/adapta-se abaixo – Figura 54 – o CDSTCU em formato genérico e global, contemplando as principais etapas, para não inibir sua aplicação/adaptação em outros contextos, agregando outras análises.

Figura 40 – CDSTCU - Modelo genérico proposto – Macro-etapas Final



Fonte: dados de pesquisa (2022)

Na Figura 54, onde foi desenhado o modelo em suas macro-etapas, nota-se um resumo gráfico de cada parte componente de cada um dos ciclos de Captura, Entendimento e Apresentação dos dados, já comentados ao longo deste trabalho,

adicionando-se, ao final, a etapa de *ACT* ou Ação/Agir, direcionada aos gestores públicos, para que ajam para melhoria das cidades, tendo em vista os dados obtidos.

Interessante notar, também, a identificação de tópicos por meio da LDA, representada, na figura, pelas figuras geométricas coloridas, sendo que, havendo a identificação do (ou dos) tópico de interesse, a coleta posterior é feita de maneira focalizada, levando em conta este tópico escolhido, que está representado pelo círculo vermelho.

Uma vez mais argumenta-se que este modo de apresentação mais genérico não exclui a importância das etapas explicadas no fluxo completo, mas tem o intuito de permitir a aplicação do CDSTCU tomando por base seu *core* de elementos-chave, adaptando-se conforme necessário.

Tendo sido concluída esta seção de Resultados, passa-se, na sequência à sessão de Conclusão, onde também são tecidas limitações e oportunidades para trabalhos futuros.

5 CONCLUSÃO

O presente trabalho teve por intuito a proposição de um modelo para extração e análise de dados/opinião pública, tendo por fonte a rede social Twitter, geograficamente situados no contexto das cidades intermediárias, tendo Maringá, como lócus do estudo.

Ademais, intentou a a) Elaboração de um código para extração de dados utilizando a abordagem de usuários centrais, por meio da linguagem aberta de programação Python, a Identificação do(s) tópico(s) proeminente(s), emergido(s) no conjunto de dados extraídos, a realização da extração de dados temática, na sequência, com base no(s) principal(is) tópico(s) identificado(s), utilizando a abordagem de usuários centrais, e também a abordagem geolocalizada, e, finalmente a aplicação de modelos de Análise de Sentimento, no conjunto de dados temáticos extraídos, referente ao(s) principal(is) tópico(s).

Fora elaborado o código/*script Python* que possibilitou a coleta de dados na cidade de Maringá, utilizando-se da abordagem de usuários centrais, sendo estes, centrados nas contas oficiais do prefeito e da prefeitura, da referida cidade, na rede social *Twitter*.

No primeiro teste de modelo, a coleta inicial, com o objetivo da análise por meio da modelagem de tópicos revelou que o principal tópico, no período considerado foi relacionado a questões da saúde pública, sobretudo à pandemia, bastante presente no início do ano de 2022, data em que foi realizada a extração dos dados. No segundo teste de modelo, a coleta para modelagem de tópicos indicou como tópico proeminente questões relativas à mobilidade urbana.

Foram então tomadas as principais palavras significativas dos tópicos encontrados, e foram utilizadas para realização das coletas temáticas considerando, novamente, a abordagem de usuários centrais, unida, desta vez, a abordagem geolocalizada, ambas temáticas, para os assuntos dos tópicos principais, para ambos os testes de modelo.

Para o primeiro teste, formou-se assim um *corpus/dataset* de documentos resultando no quantitativo de 372 *tweets,* compreendidos num período de 21 dias, e quanto ao segundo teste, um *corpus*/*dataset* de 581 *tweets*, num período de por votla de 3 meses. Foi realizada a análise de sentimento na sequência.

Em abmos os casos verificaram-se que a polaridade de sentimento predominante dos *netizens* sobre as temáticas foi negativa.

Observou-se, também, que a estratégia supervisionada foi mais bem sucedida para classificação da polaridade de sentimento dos *tweets*, sobretudo os métodos SVM e *Random Forest*, no primeiro teste de modelo, tendo sido os dois classificadores mantidos no modelo para realização do segundo teste de modelo. De um modo geral, o desempenho dos classificadores acabou sendo melhor no segundo teste de modelo, utilizando-se o pré-processamento. No primeiro teste de modelo, a inclusão, ou não, desta fase de limpeza apresentou diferenças marginais, apenas.

No primeiro teste de modelo, os picos de produção de *tweets,* no *dataset,* relacionados ao assunto temático de extração, foram devidos à edição de decretos sanitários, para o combate ao vírus da Covid-19, o que gerou discussão entre os *netizens* de Maringá. No segundo teste de modelo, os picos de produção de *tweets* foram resultado de manifestações frente ao resultado do segundo turno das eleições de 2022, por vezes sendo indicado bloqueios/interdições no trânsito.

Dentre os metadados/*features*, associados aos *tweets*, aquele que teve maior influência na determinação do sentimento, em abmos os testes, foi o metadado *"favourites_count"*, que indica o histórico, numérico, de *"likes"* dados pela fonte produtora do *tweet*, ao longo do período de atividade da conta. Observou-se que a maior média de "favoritações" históricas do usuário, ficou associado àqueles que produziram tweets com polaridade de sentimento negativa, o que pode indicar que usuários mais "ativos" do Twitter tendem a produzir tweets de cunho mais negativo. Todavia, argumenta-se sobre sua baixa usabilidade para gestores públicos, na medida em que não implica em dados nos quais se possa agir, objetivamente, mas requer interpretação que serve mais para um entendimento aprofundado de fatores, de certo modo "reprimidos" ou "escondidos", que possam influenciar no sentimento.

Referente às maiores contribuições resultantes: a extração situada no contexto das cidades, fazendo uso tanto da abordagem de usuários centrais, como da abordagem geolocalizada foi bem-sucedida, sobretudo levando em conta o fato que se trata de cidade brasileira de médio porte. A aplicação da LDA obteve êxito, indicando os tópicos mais discutidos, o que possibilitou a constatação das temáticas mais discutidas pelos *netizens*. A análise de sentimentos foi razoavelmente bem-sucedida, se levar em conta o tamanho dos datasets, sobretudo para classificar sentimentos neutros e negativos, a julgar pela consistência dos indicadores, ficando a

classificação de tweets positivos, principalmente no primeiro teste de modelo, como *shortcoming*.

Desde já se indica a limitação de se utilizar dados de redes sociais, tendo em vista não refletirem a opinião, amostral, de toda a população alvo do estudo, mas sim, somente daqueles que utilizam a referida rede social. Contudo, de forma alguma essa abordagem deve ser desconsiderada, até porque a tendência de utilização de redes sociais é uma crescente, e tende, a cada vez mais, no futuro, refletir a opinião de parcela maior da população objeto do estudo.

Também se comenta, como limitação, o fato de se ter realizado um teste de modelo com caso único, ou seja, com uma única cidade *lócus* de aplicação. Este é um fator que pode ter impacto nos resultados, uma vez que questões culturais podem entrar em jogo, na medida em que a utilização do *twitter* – e redes sociais, como um todo – pode se dar de maneira diferente nas diferentes regiões do país, a depender de fatores como acesso à tecnologia, por exemplo.

Tem-se, também, como limitação, o fato de o modelo considerar apenas os dados da rede social *Twitter*, pelas razões argumentas no decorrer do trabalho, o que pode acabar restringindo a abrangência “amostral” dos dados do trabalho. Contudo, tão logo outras plataformas permitam o *scraping* de dados, de maneira menos burocrática/mais intuitiva, o modelo pode ser adaptado/reestruturado para contemplar outras redes. Indica-se a limitação da API gratuita, considerando a impossibilidade de abstrair dados anteriores aos últimos 9 dias.

Reforça-se, uma vez mais, as contribuições teóricas e práticas deste estudo.

Sob o ponto de vista teórico, a construção do modelo para aplicação em cidades intermediárias, brasileiras, vem a suprir a lacuna de não haver encontrado, na literatura revisada e qualificada, nenhum trabalho que tenha tido este cenário como *locus* de estudo ou teste de modelo. Ainda, a abordagem de extração por meio de Usuários Centrais, além de ser uma forma escalável e mais eficiente, permitiu a coleta de maneira geograficamente situada, na cidade *locus* da aplicação, e pavimentou a utilização, posterior, da abordagem geolocalizada.

Ressalta-se, também, que, o modelo proposto de forma alguma impede a pesquisa temática, e nada impediria de, havendo a definição de um assunto concernente à administração pública, iniciar a execução a partir do segundo ciclo CUP, já definindo as *Keywords* do assunto em questão, a priori, contudo, reputa-se sendo de bom tom, executar o primeiro ciclo CUP, de modo a ter, de maneira

exploratória, os tópicos de interesse da população, no momento em que se está operando o modelo, até mesmo para não ignorar situações patentes que podem estar acontecendo.

Sob a ótica das implicações práticas, ratifica-se que as informações obtidas na análise podem embasar tomada de decisão por parte dos gestores públicos, identificando os assuntos mais comentados pelos cidadãos, prensentes na rede social, e verificando sua percepção a respeito destes. É possível, também, entender as informações das análises como um indicador da efetividade da execução de políticas públicas, partindo do contentamento ou não das pessoas das ações executadas pela administração, servindo, na mais branda das hipóteses, enquanto ferramenta complementar e mais uma fonte de dados a ser levada em consideração para uso na gestão.

Por exemplo, o nível de descontentamento da população perante a pandemia, foi agravada diante da edição de dois decretos, em datas distintas, que previam aplicação de multas para o não uso de máscaras em ambientes abertos, gerando sentimento negativo nas pessoas, com picos de *tweets*, como pôde ser notado nos resultados. Com a aplicação do modelo, a administração poderia, tempestivamente, rever ação referida, e estudar novas formas de conter o avanço da pandemia, sem tão grande recrudescimento do sentimento negativo, pela população.

Ainda sobre o aspecto prático, reforça-se o elo entre a usabilidade do modelo, e os conceitos de cidades inteligentes, na medida em que o CDSTCU se traduz enquanto um novo mecanismo que possibilita a verificação dos principais anseios dos cidadãos, de maneira exploratória, utilizando de ferramentas da TIC para aproximar o poder público de seus governados. Em última análise, o modelo possibilita aos tomadores de decisão, façam uso de mais uma fonte de informação, tanto de maneira isolada, como em apoio a outras fontes de dados já existentes, de modo a serem mais assertivos em suas ações, em prol de prover melhores serviços à população, considerando um contexto de cidades inteligentes.

Como sugestões para trabalhos futuros, inclusive para superação de limitações relatadas neste: Treinar os modelos de classificação com conjuntos de dados já testados, maiores, de modo a aumentar a consistência do desempenho nos indicadores, sobretudo para tweets positivos. Não foi encontrado trabalho na literatura revisada que utilizasse *tweets* em português, assim, a construção do referido dataset é uma oportunidade para pesquisas futuras.

Ainda como sugestões para pesquisas futuras: considera-se poder ser realizado o uso do CDSTCU, partindo do segundo ciclo CUP, se já houver um tema pré-definido, de interesse da administração pública. Superadas as limitações da twitter API, entende-se valorosa a tentativa de adaptação do modelo para rodar em tempo real, em prol da tempestividade das análises. Ainda, recomenda-se a aplicação do modelo em cidades maiores, visto que, a sustentabilidade da abordagem aqui proposta pode ser beneficial, mesmo considerando a maior facilidade e disponibilidade de acesso à *datasets* maiores. Por fim, de modo a expandir o estudo do objeto, recomenda-se que trabalhos futuros verifiquem a percepção dos gestores públicos frente à usabilidade do modelo, e aspectos que podem ser melhorados para aplicação prática.

REFERÊNCIAS

ABDUL-RAHMAN, Mohammed et al. A framework to simplify pre-processing location-based social media big data for sustainable urban planning and management. **Cities**, v. 109, p. 102986, 2021.

ADAMU, Hassan et al. Framing twitter public sentiment on Nigerian government COVID-19 palliatives distribution using machine learning. **Sustainability**, v. 13, n. 6, p. 3497, 2021.

ALIZADEH, Tooran; SARKAR, Somwrita; BURGOYNE, Sandy. Capturing citizen voice online: Enabling smart participatory local government. **Cities**, v. 95, p. 102400, 2019.

ALKHATIB, Manar et al. A sentiment reporting framework for major city events: Case study on the China-United States trade war. **Journal of Cleaner Production**, v. 264, p. 121426, 2020.

ALKHATIB, Manar; EL BARACHI, May; SHAALAN, Khaled. An Arabic social media based framework for incidents and events monitoring in smart cities. **Journal of Cleaner Production**, v. 220, p. 771-785, 2019.

ALSAEDI, Nasser; BURNAP, Pete; RANA, Omer. Can we predict a riot? Disruptive event detection using Twitter. **ACM Transactions on Internet Technology (TOIT)**, v. 17, n. 2, p. 1-26, 2017.

ANDREWS, Simon et al. Creating corroborated crisis reports from social media data through formal concept analysis. **Journal of Intelligent Information Systems**, v. 47, n. 2, p. 287-312, 2016.

ASMUSSEN, Claus Boye; MØLLER, Charles. Smart literature review: a practical topic modelling approach to exploratory literature review. **Journal of Big Data**, v. 6, n. 1, p. 1-18, 2019.

BATTY, Michael et al. Smart cities of the future. **The European Physical Journal Special Topics**, v. 214, n. 1, p. 481-518, 2012.

BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993-1022, 2003.

BOYD, Danah M.; ELLISON, Nicole B. Social network sites: Definition, history, and scholarship. **Journal of computer-mediated Communication**, v. 13, n. 1, p. 210-230, 2007.

CARAGLIU, Andrea; DEL BO, Chiara; NIJKAMP, Peter. **Smart cities in Europe. 3rd Central European Conference in Regional Science** – CERS, 2009. p. 45-59.

CAUCHICK MIGUEL, Paulo Augusto et al. Metodologia de pesquisa em engenharia de produção e gestão de operações. **Rio de Janeiro: Elzevir**, 2012.

CHOURABI, Hafedh et al. Understanding smart cities: An integrative framework. In: **2012 45th Hawaii international conference on system sciences**. IEEE, 2012. p. 2289-2297.

COSTA, Daniel G. et al. Twittersensing: An event-based approach for wireless sensor networks optimization exploiting social media in smart city applications. **Sensors**, v. 18, n. 4, p. 1080, 2018.

CRAGLIA, Max; OSTERMANN, F.; SPINSANTI, Laura. Digital Earth from vision to practice: making sense of citizen-generated content. **International Journal of Digital Earth**, v. 5, n. 5, p. 398-416, 2012.

DE CARVALHO, Carlos Henrique Ribeiro. **Desafios da mobilidade urbana no Brasil**. Texto para discussão, 2016.

EL-DIRABY, Tamer; SHALABY, Amer; HOSSEINI, Moein. Linking social, semantic and sentiment analyses to support modeling transit customers' satisfaction: Towards formal study of opinion dynamics. **Sustainable Cities and Society**, v. 49, p. 101578, 2019.

ENSSLIN, Leonardo; ENSSLIN, Sandra Rolim; PINTO, Hugo de Moraes. Processo de investigação e Análise bibliométrica: Avaliação da Qualidade dos Serviços Bancários. **Revista de administração contemporânea**, v. 17, p. 325-349, 2013.

FAN, Weiguo; GORDON, Michael D. The power of social media analytics. **Communications of the ACM**, v. 57, n. 6, p. 74-81, 2014.

FAN, Chao; JIANG, Yucheng; MOSTAFAVI, Ali. Social sensing in disaster city digital twin: Integrated textual–visual–geo framework for situational awareness during built environment disruptions. **Journal of Management in Engineering**, v. 36, n. 3, p. 04020002, 2020.

FIRJAN. Disponível em: https://www.firjan.com.br/ifdm/consulta-ao-indice/ifdm-indice-firjan-de-desenvolvimento-municipal-resultado.htm?UF=PR&IdCidade=411520&Indicador=1&Ano=2016. Acesso em 7 de Março de 2022.

GASCO, Luis et al. Beyond sound level monitoring: Exploitation of social media to gather citizens subjective response to noise. **Science of The Total Environment**, v. 658, p. 69-79, 2019.

GIFFINGER, Rudolf et al. Smart cities–Ranking of European medium-sized cities, Centre of Regional Science, Vienna. **Final Report. www. smart-cities. eu/download/smart cities final report. pdf. Erişim tarihi**, v. 12, p. 2020, 2007.

GIL, Antonio Carlos et al. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 2002.

HASNAT, Md Mehedi; HASAN, Samiul. Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data. **Transportation Research Part C: Emerging Technologies**, v. 96, p. 38-54, 2018.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101728_folder.pdf. Acesso em 27 de Janeiro de 2022.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. Disponível em: https://cidades.ibge.gov.br/brasil/pr/maringa/panorama. Acesso em 27 de Janeiro de 2022.

JAIN, Vinay Kumar; KUMAR, Shishir. Effective surveillance and predictive mapping of mosquito-borne diseases using social media. **Journal of Computational Science**, v. 25, p. 406-415, 2018.

JAMES, Gareth et al. **An introduction to statistical learning**. New York: springer, 2013.

JOSEPH, Nimish et al. Review of discussions on internet of things (IoT): insights from twitter analytics. **Journal of Global Information Management (JGIM)**, v. 25, n. 2, p. 38-51, 2017.

KANKANAMGE, Nayomi et al. Determining disaster severity through social media analysis: Testing the methodology with South East Queensland Flood tweets. **International journal of disaster risk reduction**, v. 42, p. 101360, 2020.

KAZMAIER, Jacqueline; VAN VUUREN, Jan H. A generic framework for sentiment analysis: Leveraging opinion-bearing data to inform decision making. **Decision Support Systems**, v. 135, p. 113304, 2020.

LAZAROIU, George Cristian; ROSCIA, Mariacristina. Definition methodology for the smart cities model. **Energy**, v. 47, n. 1, p. 326-332, 2012.

LI, Youzhu et al. What causes different sentiment classification on social network services? evidence from weibo with genetically modified food in China. **Sustainability**, v. 12, n. 4, p. 1345, 2020.

LIMA, Edson Pinheiro de; COSTA, Sérgio Eduardo Gouvêa da. Uma metodologia para a condução do processo associado ao projeto organizacional de sistemas de operações integradas. **Production**, v. 14, p. 18-35, 2004.

LIU, Shaojie; TENG, Jing; GONG, Yue. Extraction method and integration framework for perception features of public opinion in transportation. **Sustainability**, v. 13, n. 1, p. 254, 2021.

LU, Hao et al. Using adverse weather data in social media to assist with city-level traffic situation awareness and alerting. **Applied Sciences**, v. 8, n. 7, p. 1193, 2018.

MCARDLE, Gavin et al. Using digital footprints for a city-scale traffic simulation. **ACM Transactions on Intelligent Systems and Technology (TIST)**, v. 5, n. 3, p. 1-16, 2014.

MENDONÇA, Marcel et al. Improving public safety at fingertips: A smart city experience. In: **2016 IEEE International Smart Cities Conference (ISC2).** IEEE, 2016. p. 1-6.

MINAYO, Maria Cecília de Souza. Pesquisa social: teoria, método e criatividade. In: **Pesquisa social: teoria, método e criatividade**. 2002.

MORA, Higinio et al. Analysis of social networking service data for smart urban planning. **Sustainability**, v. 10, n. 12, p. 4732, 2018.

MUSTO, Cataldo et al. CrowdPulse: A framework for real-time semantic analysis of social streams. **Information Systems**, v. 54, p. 127-146, 2015.

OAD, Ammar et al. VADER Sentiment Analysis without and with English Punctuation Marks. In: **2021 International Journal of Advanced Trends in Computer Science and Engineering**. Warse v. 10. v.2.

POORAZIZI, Mohammad Ebrahim; HUNTER, Andrew JS; STEINIGER, Stefan. A volunteered geographic information framework to enable bottom-up disaster management platforms. **ISPRS International Journal of Geo-Information**, v. 4, n. 3, p. 1389-1422, 2015.

PREFEITURA DE MARINGÁ [Site Institucional] Disponível em: http://www2.maringa.pr.gov.br/turismo/?cod=nossa-cidade/2. Acesso em 27 de Janeiro de 2022.

PREFEITURA DE MARINGÁ [Site Institucional]. Disponível em: http://www2.maringa.pr.gov.br/cdn-imprensa/DECRETO19-2022.pdf. Acesso em 9 de Fevereiro de 2022.

PREFEITURA DE MARINGÁ [Site Institucional]. Disponível em: http://www2.maringa.pr.gov.br/sistema/arquivos/9f26c0431561.pdf. Acesso em 9 de Fevereiro de 2022.

PREFEITURA DE MARINGÁ [Site Institucional]. Disponível em: http://www2.maringa.pr.gov.br/site/noticias/2021/09/02/ranking-aponta-maringa-como-a-cidade-mais-inteligente-do-pr-com-populacao-entre-100-mil-e-500-mil-habitantes/38331. Acesso em 17 de Janeiro de 2022.

PREFEITURA DE MARINGÁ [Site Institucional]. Disponível em: http://www2.maringa.pr.gov.br/turismo/?cod=prefeitos. Acesso em 1 de Março de 2022.

RAHIMI-GOLKHANDAN, Armin; GARVIN, Michael J.; WANG, Qi. Assessing the impact of transportation diversity on postdisaster intraurban mobility. **Journal of Management in Engineering**, v. 37, n. 1, p. 04020106, 2021.

RICHARDSON, Roberto Jarry et al. **Pesquisa social: métodos e técnicas**. São Paulo: Atlas, 1999.

RODRIGUES, Ana Lúcia. Características do processo de urbanização de Maringá, PR: uma cidade de “porte médio”. **Cadernos Metrópole**, n. 12, 2004.

RUDIO, Franz Victor. Introdução ao projeto de pesquisa científica. ed.34. Petrópolis, Vozes, 2007.

SAKURAI, Mihoko; ADU-GYAMFI, Bismark. Disaster-resilient communication ecosystem in an inclusive society–A case of foreigners in Japan. **International journal of disaster risk reduction**, v. 51, p. 101804, 2020

SÁNCHEZ-ÁVILA, Mario et al. Detection of Barriers to Mobility in the Smart City Using Twitter. **IEEE Access**, v. 8, p. 168429-168438, 2020.

SAVI, Elise, CORDOVIL, Fabíola Castelo de Souza. Organização social do território e mobilidade urbana em maringá. *In*: RODRIGUES, Ana Lúcia. **Metrópoles: Território, Coesão Social e Governança Democrática, Maringá: transformações na ordem urbana**. 2015. p.307-333.

SDOUKOPOULOS, ALEXANDROS et al. Use of Social Media for Assessing Sustainable Urban Mobility Indicators. **Int. J. Sus. Dev. Plann**. v. 13, n. 2, p. 338–348, 2018.

SIEVERT, Carson; SHIRLEY, Kenneth. LDAvis: A method for visualizing and interpreting topics. In: **Proceedings of the workshop on interactive language learning, visualization, and interfaces**. 2014. p. 63-70.

SILVA, Andresa Lourenço da. Breve discussão sobre o conceito de cidade média. **Geoingá: Revista do Programa de Pós-Graduação em Geografia**, v. 5, n. 1, p. 58-76, 2013.

TSE, Rita et al. Social network based crowd sensing for intelligent transportation and climate applications. **Mobile Networks and Applications**, v. 23, n. 1, p. 177-183, 2018.

TWITTER INC. Disponível em: https://developer.twitter.com/en/docs/counting-characters. Acesso em 5 de Fevereiro de 2022.

TWITTER INC. Disponível em: https://developer.twitter.com/en/docs/twitter-api. Acesso em 5 de Fevereiro de 2022.

ZHANG, Lu Fan Zhang et al. Social Network Based Crowd Sensing for Intelligent Transportation and Climate Applications. 2017.

UNITED NATIONS, DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS, POPULATION DIVISION. World urbanization prospects 2018. **Population Division**, 2018.

UNITED NATIONS. Disponível em: https://www.br.undp.org/content/brazil/pt/home/idh0/rankings/idhm-municipios-2010.html. Acesso em 1 de Março de 2022.

UNITED NATIONS. Disponível em: https://www.un.org/development/desa/dpad/publication/world-economic-situation-and-prospects-february-2020-briefing-no-134/. Acesso em 27 de Janeiro de 2022.

URBAN SYSTEMS. Disponível em: https://ranking.connectedsmartcities.com.br/. Acesso em 1 de Março de 2022.

URBAN SYSTEMS. Disponível em: https://ranking.connectedsmartcities.com.br/. Acesso em 13 de Janeiro de 2023.

USHARANI, B. Analysis of Supervised and Unsupervised Learning Classifiers for Online Sentiment Analysis. *In*: 2018 **Asian Journal of Computer Science Engineering 2018**. p.17-21.

WANG, Yandong et al. Using social media for emergency response and urban sustainability: A case study of the 2012 Beijing rainstorm. **Sustainability**, v. 8, n. 1, p. 25, 2016.

WANG, Ruo-Qian et al. Tracking flooding phase transitions and establishing a passive hotline with ai-enabled social media data. **IEEE Access**, v. 8, p. 103395-103404, 2020.

WICKELMAIER, Florian. An introduction to MDS. **Sound Quality Research Unit, Aalborg University, Denmark**, v. 46, n. 5, p. 1-26, 2003.

YUAN, Faxi et al. Internet of people enabled framework for evaluating performance loss and resilience of urban critical infrastructures. **Safety science**, v. 134, p. 105079, 2021.

YUAN, Yihong et al. The missing parts from social media–enabled smart cities: Who, where, when, and what?. **Annals of the American Association of Geographers**, v. 110, n. 2, p. 462-475, 2020.

APÊNDICE A – *Scripts* de extração de dados .py

CUP1 – Extração Ñ temática (*extração sem assunto/domínio prévio*)

```
#IMPORTANDO PACKAGES
import tweepy
import pandas as pd

#AUTENTICAÇÃO
auth = tweepy.OAuthHandler("consumer_key", "consumer_secret")
auth.set_access_token("key", "secret")
api = tweepy.API(auth)

#LISTA QUE SERÁ APENDIDA
tweets_list = []
source_accounts = "@prefeiturademga OR @UlissesMaia"

#APENDAR TODOS OS CONTEÚDOS DOS TWEETS NA TWEETS_LIST COM CONTADOR
count = 0
for tweet in tweepy.Cursor(api.search_tweets, q=f"{source_accounts} -filter:retweets -
from:prefeiturademga -from:UlissesMaia", tweet_mode="extended").items(1000):
    tweets_list.append(tweet)
    tweets_list.append(tweet.user) #as informações do usuário criador do tweet ficam na linha
imediatamente abaixo a das infos do tweet
    count += 1

    #CONTROLE DE VISUALIZAÇÃO NA PYTHON CONSOLE
    print("texto do tweet..: ", tweet.full_text, "\n",
          "data e hora do tweet..: ", tweet.created_at, "\n",
          "o id do usuário que tuitou..: ", tweet.id_str, "\n",
          "@, no twitter, do usuário que tuitou..: ", tweet.user.screen_name, "\n",
          "nome do usuário que tuitou..: ", tweet.user.name, "\n",
          "o id do usuário ao qual foi dirigido o tweet..: ", tweet.in_reply_to_user_id_str, "\n",
          "o nome do usuário ao qual foi dirigido o tweet..: ", tweet.in_reply_to_screen_name, "\n",
          "localidade, caso haja, do usuário que tuitou..: ", tweet.user.location, "\n",
          "descrição, caso haja, do perfil do usuário que tuitou..: ", tweet.user.description, "\n",
          "número de seguidores do usuário que tuitou..: ", tweet.user.followers_count, "\n",
          "número de pessoas que o usuário que tuitou segue..: ", tweet.user.friends_count, "\n",
          "informação de se o perfil do usuário que tuitou é verificado..: ", tweet.user.verified, "\n",
          "informação de se o usuário que tuitou permite o compartilhamento da localização de suas
postagens..:",
          tweet.user.geo_enabled, "\n",
          "informações de se tratar ou não de tweet geotagged, de alguma forma..: ", tweet.geo, " ",
tweet.coordinates,
          " ", tweet.place, "\n",
          "número de retweets do tweet em questão..: ", tweet.retweet_count, "\n",
          "número de favoritações do tweet em questão..: ", tweet.favorite_count, "\n")

print("qtde de scraped tweets: ", count)

#CONVERTER A LISTA TWEETS_LIST PARA UM DATAFRAME PANDAS
tweets_list_DF = pd.DataFrame(vars(tweets_list[i]) for i in range(len(tweets_list)))

#DEFINIÇÃO DE ONDE SALVAR O ARQUIVO CSV COM O DATAFRAME
caminho_arquivo = "C:\\Users\\arthu\\Dropbox\\Py_Tweets\\tweets_saúde_maringá.csv"

#USAR O PANDAS PARA SALVAR O DATAFRAME PARA CSV
tweets_list_DF.to_csv(caminho_arquivo)
```

CUP2 - Pesquisa Temática-Saúde_Central Users

```
#IMPORTANDO PACKAGES
import tweepy
import pandas as pd
```

```
#AUTENTICAÇÃO
auth = tweepy.OAuthHandler("consumer_key", "consumer_secret")
auth.set_access_token("key", "secret")
api = tweepy.API(auth)

#LISTA QUE SERÁ APENDIDA
tweets_list = []
source_accounts = "@UlissesMaia OR @prefeiturademga"
keywords_saude = """saude OR covid OR covid-19 OR UPA OR máscara OR máscaras OR
isolamento OR isolado OR médico OR
médicos OR infectologista OR infectologistas OR teste OR testes OR pandemia OR pandemico OR
vacinação OR vacina OR
vacinado OR vacinados OR vacinar OR posto OR postinho OR passaporte OR reforço"""
contas_filtradas = "-from:prefeiturademga -from:UlissesMaia -from:cbnmaringa -from:Maringa_Post
-from:diario_maringa " \
                "-from:MaringaCom - from:portalgmconline"


#APENDAR TODOS OS CONTEÚDOS DOS TWEETS NA TWEETS_LIST COM CONTADOR
count = 0
for tweet in tweepy.Cursor(api.search_tweets,q=f"{source_accounts} AND {keywords_saude} AND
{contas_filtradas} -filter:retweets",
                        tweet_mode="extended").items(1000):
    tweets_list.append(tweet)
    tweets_list.append(tweet.user) #as informações do usuário criador do tweet ficam na linha
imediatamente abaixo a das infos do tweet
    count += 1

    #CONTROLE DE VISUALIZAÇÃO NA PYTHON CONSOLE
    print("texto do tweet..: ", tweet.full_text, "\n",
          "data e hora do tweet..: ", tweet.created_at, "\n",
          "o id do usuário que tuitou..: ", tweet.id_str, "\n",
          "@, no twitter, do usuário que tuitou..: ", tweet.user.screen_name, "\n",
          "nome do usuário que tuitou..: ", tweet.user.name, "\n",
          "o id do usuário ao qual foi dirigido o tweet..: ", tweet.in_reply_to_user_id_str, "\n",
          "o nome do usuário ao qual foi dirigido o tweet..: ", tweet.in_reply_to_screen_name, "\n",
          "localidade, caso haja, do usuário que tuitou..: ", tweet.user.location, "\n",
          "descrição, caso haja, do perfil do usuário que tuitou..: ", tweet.user.description, "\n",
          "número de seguidores do usuário que tuitou..: ", tweet.user.followers_count, "\n",
          "número de pessoas que o usuário que tuitou segue..: ", tweet.user.friends_count, "\n",
          "informação de se o perfil do usuário que tuitou é verificado..: ", tweet.user.verified, "\n",
          "informação de se o usuário que tuitou permite o compartilhamento da localização de suas
postagens..:",
          tweet.user.geo_enabled, "\n",
          "informações de se tratar ou não de tweet geotagged, de alguma forma..: ", tweet.geo, " ",
tweet.coordinates,
          " ", tweet.place, "\n",
          "número de retweets do tweet em questão..: ", tweet.retweet_count, "\n",
          "número de favoritações do tweet em questão..: ", tweet.favorite_count, "\n")

print("qtde de scraped tweets: ", count)

#CONVERTER A LISTA TWEETS_LIST PARA UM DATAFRAME PANDAS
tweets_list_DF = pd.DataFrame(vars(tweets_list[i]) for i in range(len(tweets_list)))

#DEFINIÇÃO DE ONDE SALVAR O ARQUIVO CSV COM O DATAFRAME
caminho_arquivo = "C:\\Users\\arthu\\Dropbox\\Py_Tweets\\Pesquisa
Temática\\tweets_saúde__maringá_central_accounts.csv"
```

```
#USAR O PANDAS PARA SALVAR O DATAFRAME PARA CSV
tweets_list_DF.to_csv(caminho_arquivo)
```

CUP2 - Pesquisa Temática-Saúde_Geolocalizada

```
#IMPORTANDO PACKAGES
import tweepy
import pandas as pd

#AUTENTICAÇÃO
auth = tweepy.OAuthHandler("consumer_key", "consumer_secret")
auth.set_access_token("key", "secret")
api = tweepy.API(auth)

#LISTA QUE SERÁ APENDIDA
tweets_list = []
source_accounts = "@prefeiturademga OR @UlissesMaia"
menção_cidade = "Maringá OR cidade OR UlissesMaia OR Maia OR prefeito OR prefeitura"
keywords_saude = """saude OR covid OR covid-19 OR UPA OR máscara OR máscaras OR
isolamento OR isolado OR médico OR
médicos OR infectologista OR infectologistas OR teste OR testes OR pandemia OR pandemico
OR vacinação OR vacina OR
vacinado OR vacinados OR vacinar OR posto OR postinho OR passaporte OR reforço"""
contas_filtradas = "-from:prefeiturademga -from:UlissesMaia -from:cbnmaringa -from:Maringa_Post
-from:diario_maringa " \
                "-from:MaringaCom - from:portalgmconline"

#APENDAR TODOS OS CONTEÚDOS DOS TWEETS NA TWEETS_LIST COM CONTADOR
count = 0
for tweet in tweepy.Cursor(api.search_tweets, q=f"{keywords_saude} AND {menção_cidade} AND
{contas_filtradas} -filter:retweets",
                        geocode="-23.421031,-51.937609,10km",
                        tweet_mode="extended"
                        ).items(1000):
    tweets_list.append(tweet)
    tweets_list.append(tweet.user) #as informações do usuário criador do tweet ficam na linha
imediatamente abaixo a das infos do tweet
    count += 1

    #CONTROLE DE VISUALIZAÇÃO NA PYTHON CONSOLE
    print("texto do tweet..: ", tweet.full_text, "\n",
        "data e hora do tweet..: ", tweet.created_at, "\n",
```

```
        "o id do usuário que tuitou..: ", tweet.id_str, "\n",
        "@, no twitter, do usuário que tuitou..: ", tweet.user.screen_name, "\n",
        "nome do usuário que tuitou..: ", tweet.user.name, "\n",
        "o id do usuário ao qual foi dirigido o tweet..: ", tweet.in_reply_to_user_id_str, "\n",
        "o nome do usuário ao qual foi dirigido o tweet..: ", tweet.in_reply_to_screen_name, "\n",
        "localidade, caso haja, do usuário que tuitou..: ", tweet.user.location, "\n",
        "descrição, caso haja, do perfil do usuário que tuitou..: ", tweet.user.description, "\n",
        "número de seguidores do usuário que tuitou..: ", tweet.user.followers_count, "\n",
        "número de pessoas que o usuário que tuitou segue..: ", tweet.user.friends_count, "\n",
        "informação de se o perfil do usuário que tuitou é verificado..: ", tweet.user.verified, "\n",
        "informação de se o usuário que tuitou permite o compartilhamento da localização de suas
postagens..:",
        tweet.user.geo_enabled, "\n",
        "informações de se tratar ou não de tweet geotagged, de alguma forma..: ", tweet.geo, " ",
tweet.coordinates,
        " ", tweet.place, "\n",
        "número de retweets do tweet em questão..: ", tweet.retweet_count, "\n",
        "número de favoritações do tweet em questão..: ", tweet.favorite_count, "\n")

print("qtde de scraped tweets: ", count)

#CONVERTER A LISTA TWEETS_LIST PARA UM DATAFRAME PANDAS
tweets_list_DF = pd.DataFrame(vars(tweets_list[i]) for i in range(len(tweets_list)))

#DEFINIÇÃO DE ONDE SALVAR O ARQUIVO CSV COM O DATAFRAME
caminho_arquivo = "C:\\Users\\arthu\\Dropbox\\Py_Tweets\\Pesquisa
Temática\\geo_tweets_saúde__maringá.csv"

#USAR O PANDAS PARA SALVAR O DATAFRAME PARA CSV
tweets_list_DF.to_csv(caminho_arquivo)
```

CUP2 - Pesquisa Temática-Mobilidade_Central Users

```
#IMPORTANDO PACKAGES
import tweepy
import pandas as pd

#AUTENTICAÇÃO
auth = tweepy.OAuthHandler("consumer_key", "consumer_secret")
auth.set_access_token("key", "secret")
api = tweepy.API(auth)

#LISTA QUE SERÁ APENDIDA
tweets_list = []
source_accounts = "@UlissesMaia OR @prefeiturademga"
            keywords_mobilidade = """ônibus OR moto OR motoqueiro OR caminhão OR carro OR
rua OR avenida OR multa OR transporte OR cruzar OR cruzamento OR vaga OR ciclo OR pista
OR rodovia OR sinal OR semáforo OR acidente OR buraco OR mobilidade OR contorno OR
```

```
viaduto OR passagem OR passageiro OR estacionar OR estacionamento OR aslfato OR aslfaltar
OR calçada OR pedestre OR pedestres OR trafego OR trânsito"""
contas_filtradas = "-from:prefeiturademga -from:UlissesMaia -from:cbnmaringa -from:Maringa_Post
-from:diario_maringa " \
                "-from:MaringaCom - from:portalgmconline"


#APENDAR TODOS OS CONTEÚDOS DOS TWEETS NA TWEETS_LIST COM CONTADOR
count = 0
for tweet in tweepy.Cursor(api.search_tweets,q=f"{source_accounts} AND {keywords_mobilidade}
AND {contas_filtradas} -filter:retweets",
                        tweet_mode="extended").items(1000):
    tweets_list.append(tweet)
    tweets_list.append(tweet.user) #as informações do usuário criador do tweet ficam na linha
imediatamente abaixo a das infos do tweet
    count += 1

    #CONTROLE DE VISUALIZAÇÃO NA PYTHON CONSOLE
    print("texto do tweet..: ", tweet.full_text, "\n",
          "data e hora do tweet..: ", tweet.created_at, "\n",
          "o id do usuário que tuitou..: ", tweet.id_str, "\n",
          "@, no twitter, do usuário que tuitou..: ", tweet.user.screen_name, "\n",
          "nome do usuário que tuitou..: ", tweet.user.name, "\n",
          "o id do usuário ao qual foi dirigido o tweet..: ", tweet.in_reply_to_user_id_str, "\n",
          "o nome do usuário ao qual foi dirigido o tweet..: ", tweet.in_reply_to_screen_name, "\n",
          "localidade, caso haja, do usuário que tuitou..: ", tweet.user.location, "\n",
          "descrição, caso haja, do perfil do usuário que tuitou..: ", tweet.user.description, "\n",
          "número de seguidores do usuário que tuitou..: ", tweet.user.followers_count, "\n",
          "número de pessoas que o usuário que tuitou segue..: ", tweet.user.friends_count, "\n",
          "informação de se o perfil do usuário que tuitou é verificado..: ", tweet.user.verified, "\n",
          "informação de se o usuário que tuitou permite o compartilhamento da localização de suas
postagens..:",
          tweet.user.geo_enabled, "\n",
          "informações de se tratar ou não de tweet geotagged, de alguma forma..: ", tweet.geo, " ",
tweet.coordinates,
          " ", tweet.place, "\n",
          "número de retweets do tweet em questão..: ", tweet.retweet_count, "\n",
          "número de favoritações do tweet em questão..: ", tweet.favorite_count, "\n")

print("qtde de scraped tweets: ", count)

#CONVERTER A LISTA TWEETS_LIST PARA UM DATAFRAME PANDAS
tweets_list_DF = pd.DataFrame(vars(tweets_list[i]) for i in range(len(tweets_list)))

#DEFINIÇÃO DE ONDE SALVAR O ARQUIVO CSV COM O DATAFRAME
caminho_arquivo = "C:\\Users\\arthu\\Dropbox\\Py_Tweets\\Pesquisa
Temática\\tweets_saúde__maringá_central_accounts.csv"

#USAR O PANDAS PARA SALVAR O DATAFRAME PARA CSV
tweets_list_DF.to_csv(caminho_arquivo)
```

CUP2 - Pesquisa Temática-Mobilidade_Geolocalizada1

```
#IMPORTANDO PACKAGES

import tweepy

import pandas as pd


#AUTENTICAÇÃO
```

```
auth = tweepy.OAuthHandler("consumer_key", "consumer_secret")
auth.set_access_token("key", "secret")
api = tweepy.API(auth)

#LISTA QUE SERÁ APENDIDA
tweets_list = []
source_accounts = "@prefeiturademga OR @UlissesMaia"
menção_cidade = "Maringá OR cidade OR UlissesMaia OR Maia OR prefeito OR prefeitura"
keywords_mobilidade = """ônibus OR moto OR motoqueiro OR caminhão OR carro OR rua OR
avenida OR multa OR transporte OR cruzar OR cruzamento OR vaga OR ciclo OR pista OR
rodovia OR sinal OR semáforo"""
contas_filtradas = "-from:prefeiturademga -from:UlissesMaia -from:cbnmaringa -from:Maringa_Post
-from:diario_maringa " \
                "-from:MaringaCom - from:portalgmconline"

#APENDAR TODOS OS CONTEÚDOS DOS TWEETS NA TWEETS_LIST COM CONTADOR
count = 0
for tweet in tweepy.Cursor(api.search_tweets, q=f"{keywords_mobilidade} AND {menção_cidade}
AND {contas_filtradas} -filter:retweets",
                        geocode="-23.421031,-51.937609,10km",
                        tweet_mode="extended"
                        ).items(1000):
    tweets_list.append(tweet)
    tweets_list.append(tweet.user) #as informações do usuário criador do tweet ficam na linha
imediatamente abaixo a das infos do tweet
    count += 1

    #CONTROLE DE VISUALIZAÇÃO NA PYTHON CONSOLE
    print("texto do tweet..: ", tweet.full_text, "\n",
        "data e hora do tweet..: ", tweet.created_at, "\n",
        "o id do usuário que tuitou..: ", tweet.id_str, "\n",
        "@, no twitter, do usuário que tuitou..: ", tweet.user.screen_name, "\n",
        "nome do usuário que tuitou..: ", tweet.user.name, "\n",
        "o id do usuário ao qual foi dirigido o tweet..: ", tweet.in_reply_to_user_id_str, "\n",
        "o nome do usuário ao qual foi dirigido o tweet..: ", tweet.in_reply_to_screen_name, "\n",
        "localidade, caso haja, do usuário que tuitou..: ", tweet.user.location, "\n",
        "descrição, caso haja, do perfil do usuário que tuitou..: ", tweet.user.description, "\n",
        "número de seguidores do usuário que tuitou..: ", tweet.user.followers_count, "\n",
        "número de pessoas que o usuário que tuitou segue..: ", tweet.user.friends_count, "\n",
        "informação de se o perfil do usuário que tuitou é verificado..: ", tweet.user.verified, "\n",
```

```
        "informação de se o usuário que tuitou permite o compartilhamento da localização de suas
postagens..:",
        tweet.user.geo_enabled, "\n",
        "informações de se tratar ou não de tweet geotagged, de alguma forma..: ", tweet.geo, " ",
tweet.coordinates,
        " ", tweet.place, "\n",
        "número de retweets do tweet em questão..: ", tweet.retweet_count, "\n",
        "número de favoritações do tweet em questão..: ", tweet.favorite_count, "\n")

print("qtde de scraped tweets: ", count)

#CONVERTER A LISTA TWEETS_LIST PARA UM DATAFRAME PANDAS
tweets_list_DF = pd.DataFrame(vars(tweets_list[i]) for i in range(len(tweets_list)))

#DEFINIÇÃO DE ONDE SALVAR O ARQUIVO CSV COM O DATAFRAME
caminho_arquivo = "C:\\Users\\arthu\\Dropbox\\Py_Tweets\\Pesquisa
Temática\\geo_tweets_saúde__maringá.csv"

#USAR O PANDAS PARA SALVAR O DATAFRAME PARA CSV
tweets_list_DF.to_csv(caminho_arquivo)
```

CUP2 - Pesquisa Temática-Mobilidade_Geolocalizada2

```
#IMPORTANDO PACKAGES
import tweepy
import pandas as pd

#AUTENTICAÇÃO
auth = tweepy.OAuthHandler("consumer_key", "consumer_secret")
auth.set_access_token("key", "secret")
api = tweepy.API(auth)

#LISTA QUE SERÁ APENDIDA
tweets_list = []
source_accounts = "@prefeiturademga OR @UlissesMaia"
menção_cidade = "Maringá OR cidade OR UlissesMaia OR Maia OR prefeito OR prefeitura"
keywords_mobilidade = """acidente OR buraco OR mobilidade OR contorno OR viaduto OR
passagem OR passageiro OR estacionar OR estacionamento OR aslfato OR aslfaltar OR calçada
OR pedestre OR pedestres OR trafego OR trânsito"""
contas_filtradas = "-from:prefeiturademga -from:UlissesMaia -from:cbnmaringa -from:Maringa_Post
-from:diario_maringa " \
                "-from:MaringaCom - from:portalgmconline"
```

```
#APENDAR TODOS OS CONTEÚDOS DOS TWEETS NA TWEETS_LIST COM CONTADOR
count = 0
for tweet in tweepy.Cursor(api.search_tweets, q=f"{keywords_mobilidade} AND {menção_cidade} AND {contas_filtradas} -filter:retweets",
                           geocode="-23.421031,-51.937609,10km",
                           tweet_mode="extended"
                           ).items(1000):
    tweets_list.append(tweet)
    tweets_list.append(tweet.user) #as informações do usuário criador do tweet ficam na linha imediatamente abaixo a das infos do tweet
    count += 1

    #CONTROLE DE VISUALIZAÇÃO NA PYTHON CONSOLE
    print("texto do tweet..: ", tweet.full_text, "\n",
          "data e hora do tweet..: ", tweet.created_at, "\n",
          "o id do usuário que tuitou..: ", tweet.id_str, "\n",
          "@, no twitter, do usuário que tuitou..: ", tweet.user.screen_name, "\n",
          "nome do usuário que tuitou..: ", tweet.user.name, "\n",
          "o id do usuário ao qual foi dirigido o tweet..: ", tweet.in_reply_to_user_id_str, "\n",
          "o nome do usuário ao qual foi dirigido o tweet..: ", tweet.in_reply_to_screen_name, "\n",
          "localidade, caso haja, do usuário que tuitou..: ", tweet.user.location, "\n",
          "descrição, caso haja, do perfil do usuário que tuitou..: ", tweet.user.description, "\n",
          "número de seguidores do usuário que tuitou..: ", tweet.user.followers_count, "\n",
          "número de pessoas que o usuário que tuitou segue..: ", tweet.user.friends_count, "\n",
          "informação de se o perfil do usuário que tuitou é verificado..: ", tweet.user.verified, "\n",
          "informação de se o usuário que tuitou permite o compartilhamento da localização de suas postagens..:",
          tweet.user.geo_enabled, "\n",
          "informações de se tratar ou não de tweet geotagged, de alguma forma..: ", tweet.geo, " ", tweet.coordinates,
          " ", tweet.place, "\n",
          "número de retweets do tweet em questão..: ", tweet.retweet_count, "\n",
          "número de favoritações do tweet em questão..: ", tweet.favorite_count, "\n")

print("qtde de scraped tweets: ", count)

#CONVERTER A LISTA TWEETS_LIST PARA UM DATAFRAME PANDAS
tweets_list_DF = pd.DataFrame(vars(tweets_list[i]) for i in range(len(tweets_list)))

#DEFINIÇÃO DE ONDE SALVAR O ARQUIVO CSV COM O DATAFRAME
```

```
caminho_arquivo = "C:\\Users\\arthu\\Dropbox\\Py_Tweets\\Pesquisa
Temática\\geo_tweets_saúde__maringá.csv"

#USAR O PANDAS PARA SALVAR O DATAFRAME PARA CSV
tweets_list_DF.to_csv(caminho_arquivo)
```