UNIVERSIDADE DO VALE DO RIO DOS SINOS
UNIDADE ACADÊMICA DE GRADUAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

EDUARDO SIPP

UMA PROPOSTA DE MODELO COMBINANDO THICK DATA E BIG DATA PARA POTENCIALIZAR OS RESULTADOS DAS ANÁLISES DE DADOS

São Leopoldo
2021

EDUARDO SIPP

UMA PROPOSTA DE MODELO COMBINANDO THICK DATA E BIG DATA PARA POTENCIALIZAR OS RESULTADOS DAS ANÁLISES DE DADOS

Artigo apresentado como requisito parcial para obtenção do título de Bacharel em Ciência da Computação, pelo Curso de Ciência da Computação da Universidade do Vale do Rio dos Sinos (UNISINOS)

Orientador(a): Profa. Dra. Rosemary Francisco

São Leopoldo
2021

UMA PROPOSTA DE MODELO COMBINANDO THICK DATA E BIG DATA PARA POTENCIALIZAR OS RESULTADOS DAS ANÁLISES DE DADOS

Eduardo Sipp[1]
Rosemary Francisco[2]

Resumo: O big data apresenta uma revolução no que diz respeito ao armazenamento de dados. Tais dados permitem diversas análises, possibilidades para gerações de insights de valor e também para o suporte em tomadas de decisão, contudo a maioria destas análises ainda apresentam falhas ou baixa rentabilidade quando se observa o valor investido frente aos resultados obtidos. Considerando este cenário, o presente trabalho tem como objetivo principal analisar como o thick data combinado com o big data pode ser utilizado para tornar as análises mais assertivas. Em suma, o thick data busca entender o comportamento humano e como ocorre a evolução do relacionamento do ser humano com determinado produto ou serviço ao longo do tempo. Compreender tais questões permite a extração mais acurada dos dados, logo, tende a geração de melhores resultados. Para a condução da pesquisa foi utilizado o método de pesquisa Design Science Research (DSR). Ao decorrer do trabalho, é apresentada a prova de conceito do modelo desenvolvido e também duas avaliações para observar a viabilidade do mesmo, uma teórica utilizando dois cases retirados da literatura e outra prática utilizando dados obtidos através da API do Twitter e dados do Governo Federal referentes a Covid-19. O principal resultado obtido a partir deste trabalho é demonstrar a possibilidade da combinação do thick data com o big data para viabilizar análises de dados mais abrangentes. Além disso, entre as principais contribuições desta pesquisa pode-se citar o artefato gerado, isto é, o modelo que agrega o thick data à uma estrutura de big data, com foco em utilizar os dados qualitativos juntamente aos dados quantitativos.

Palavras-chave: Thick data. Big data. Análise de dados. Design Science Research.

Abstract: Big data is a revolution in terms of data storage. Such data allow for several analyzes, possibilities for generating valuable insights and to support decision-making, however, the vast majority of these analyzes still have a high failure rate or low profitability when observing the amount invested versus the results obtained. Considering this scenario, the present work has as the main objective to analyze how thick data combined with big data can be used to make the analyzes more assertive. In summary, thick data intends to understand human behavior and how the human relationship with a particular product or service evolves over time. Understanding such issues allow more accurate extraction of data, the execution of better analyzes and, consequently, tends to generate better results. To conduct the research, the Design Science Research (DSR) research method was used. The paper presents the proof of concept of the model developed as well as two evaluations to observe the viability of the model, one theoretical using two cases from the literature and another practical based on Twitter data and Federal Government data about Covid-19. The main result obtained from this work is to demonstrate the possibility of combining thick data with big data to achieve more comprehensive data analysis. Furthermore, among the main contributions of this research can be cited the artifact generated, that is, the model that adds thick data to a big data structure, with a focus on using

[1]Graduando em Ciência da Computação pela Unisinos. Email: esipp@edu.unisinos.br
[2]Docente nos cursos de TI da Unisinos. Email: rosemaryf@unisinos.br

qualitative data together with quantitative data.

Keywords: Thick data. Big data. Data analysis. Design Science Research.

1 INTRODUÇÃO

A quantidade de dados gerados cresce de forma exponencial, tanto em volume quanto em velocidade. (BORNAKKE; DUE, 2018). Aliado a isso, o mundo encontra-se em uma era onde o armazenamento de dados é barato e onde há grande interesse em coletar o máximo de dados possíveis referente ao comportamento do usuário, gerando um cenário positivo para análises que vão além de quantitativas. (SMETS; LIEVENS, 2018). Comportamento este que não se restringe a uma determinada área, pois atualmente todos os tipos de negócios e organizações estão, de alguma forma, online, deixando rastros de seus usuários ou clientes nos mais variados locais. (MOHAMMED; FIAIDHI, 2019). Diante deste cenário, conforme Mohammed e Fiaidhi (2019) para ter sucesso e crescer, é necessário que se tenha a capacidade de obter e reter esses dados com o intuito de compreender melhor o comportamento do usuário, isto é, o comportamento do ser humano em si, pois uma empresa que conhece seus clientes pode tomar decisões mais inteligentes que maximizam a lealdade do consumidor e evita gastos desnecessários.

Há muito investimento em big data atualmente, de acordo com o Gartner, 48% das organizações pesquisadas em 2016 já haviam investido em sistemas de big data e 25% pretendiam investir nos próximos dois anos. (MEULEN, 2016). Além disso, Asay (2014) também destaca o investimento que as empresas estão fazendo em cientistas de dados, com salários na média de $123.000. O que deve ser compreendido neste ponto é que não é, unicamente, o tamanho do investimento que definirá a qualidade da saída da informação. Há necessidade de um processo bem definido, assim como, o entendimento do que cada tipo de dado, quantitativo e qualitativo, pela sua natureza pode demonstrar ao ser analisado. Isso pode ser avaliado ao observar os resultados dos projetos de big data, onde encontram-se altos índices de falhas. O Gartner reportou em 2017 que 60% dos projetos de big data falharam. (MOHAMMED; FIAIDHI, 2019). Complementar a essa informação, o Gartner também informou em um estudo que das companhias que investiram em recursos de big data, apenas 8% estavam fazendo algo significativo, isto é, tendo avanços em processos importantes e não apenas avanços incrementais em processos pré-existentes ao big data. (WANG, 2016).

Então, apesar da quantidade de dados à disposição ser cada vez maior, o armazenamento ser viável e haver grandes investimentos, por que os projetos ainda tem uma taxa de sucesso tão pequena? Como bem analisaram Mohammed e Fiaidhi (2019), os projetos falharam pela obsessão em obter dados quantitativos à custa de dados qualitativos. A maioria dos projetos de análise de dados no big data focam na busca de métricas às custas de capturar os dados da perspectiva do cliente, assim apresentam dificuldades em transformar os insights dos dados em ações. (FIAIDHI, 2020). Pode-se observar que os próprios softwares na área da análise de dados, principalmente web, entregam respostas vagas quando se trata de entender o verdadeiro

valor comercial. (MOHAMMED; FIAIDHI, 2019). Isso ocorre porque o big data, por natureza, remove o contexto e não consegue trazer os traços qualitativos e etnográficos dos dados e sem esse contexto importantes decisões são tomadas com dados incompletos. (PINK et al., 2016). Um exemplo mais específico de como a análise não deve ser focada apenas em dados quantitativos é trazido por Wang (2017) onde destaca-se o grande erro que a Nokia cometeu ao compreender os sentimentos e opiniões de seus usuários, baseando-se apenas nos dados de redes sociais e outras fontes similares. Cabe o destaque para uma citação: "O que é mensurável não é o mesmo que o que é valioso". (WANG, 2017). O thick data torna-se um complemento ao big data, ajudando a entender não apenas o que as pessoas fazem e fizeram, mas também o porquê. Essa abordagem permite enriquecer o big data com insights sobre o que impulsiona as pessoas, não apenas como clientes, mas como seres humanos. (MOHAMMED; FIAIDHI, 2019).

Desta maneira, o presente estudo apresenta a seguinte problematização: é possível aliar o thick data ao big data a fim de gerar novas análises e potencializar as análises já existentes? Acredita-se que é possível utilizar dados advindos de fontes thick data para obter uma melhor compreensão dos mesmos por meio da relação de dados quantitativos e qualitativos e, por consequência, gerar análises mais assertivas. Análises mais assertivas são análises que tem um número previsto mais próximo ao número observado e, também, análises que apenas são possíveis no olhar analítico através desta nova lente. Conforme Smets e Lievens (2018) destacam, o thick data possui a capacidade de obter um entendimento mais profundo do comportamento humano dentro do seu contexto, enquanto o big data, por si só, possui seu foco voltado para a busca de padrões em grandes conjuntos de dados.

Considerando a questão da pesquisa de trabalho, o objetivo geral é analisar como o thick data pode ser aliado ao big data para possibilitar análises mais assertivas e abrangentes. Para alcançar este objetivo foram definidos os seguintes objetivos específicos:

a) identificar as técnicas de análises qualitativas que podem ser utilizadas;

b) desenvolver o modelo que possibilite a combinação de dados quantitativos (big data) com dados qualitativos (thick data);

c) avaliar o modelo com o case da Netflix e o case de cidades inteligentes referenciados na bibliografia;

d) avaliar o modelo com um case de dados advindos da API do Twitter e do Governo Federal.

O método de pesquisa escolhido para o desenvolvimento deste trabalho foi o Design Science Research (DSR). (KUECHLER; VAISHNAVI, 2008). De acordo com estes autores, o método possibilita a construção de uma ampla gama de artefatos sociotécnicos, como o modelo proposto nesta pesquisa. Além de ter como objetivo contribuir tanto para a teoria quanto para a prática, tornando-se uma abordagem promissora para esta pesquisa devido ao interesse nesta temática por parte de empresas, analistas e cientistas de dados.

Como descrito na problematização, a maioria dos projetos de big data falham e um percentual ainda maior não é rentável. Sendo assim, o thick data pode ser um diferencial de alto nível a qualquer organização, ajudando as empresas a descobrir os tipos de insights que a maioria das abordagens quantitativas não oferecem. (MOHAMMED; FIAIDHI, 2019). Além disso, grande parte dos insights do consumidor pode ser capturada com métodos de pesquisa qualitativa que funcionam com tamanhos de amostra pequenos, mesmo estatisticamente insignificantes (FIAIDHI, 2020) proporcionando competitividade a todos os tamanhos de companhias, mesmo que estas não possuem inúmeras fontes de dados ou imensas massas dos mesmos.

O presente trabalho está estruturado em cinco seções. A segunda seção apresenta a fundamentação teórica que aborda brevemente os principais conceitos que ajudam a fundamentar o problema de pesquisa, além da apresentação dos trabalhos relacionados. Após, descreve-se a metodologia utilizada. A quarta seção discorre sobre a construção do modelo proposto, seguida pela análise e discussão dos resultados. Por fim, é apresentada a conclusão.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção são apresentados os conceitos fundamentais abordados neste trabalho com base na literatura pesquisada, evidenciando a importância da utilização do thick data nas análises de dados do big data. Após, são apresentados os trabalhos relacionados.

2.1 Conceitos Fundamentais

2.1.1 Thick Data

O conceito de thick data foi popularizado por Wang (2016) que utilizou o mesmo para descrever o entendimento das peculiaridades do comportamento humano com o intuito de prever como o relacionamento de um indivíduo com determinado serviço ou produto iria evoluir ao longo do tempo. Já Mohammed e Fiaidhi (2019) destacaram a capacidade que o thick data possui de compreender os consumidores individualmente com mais precisão, oferecendo assim análises mais profundas e próximas ao histórico do consumidor. Complementar a isso, Smets e Lievens (2018) relacionam o entendimento do comportamento humano ao contexto dos dados, destacando a importância de uma das características principais do thick data: manter o contexto dos dados junto a estes. O contexto, isto é, o que foi realizado, por quem e onde, como e por que, é parte essencial para que a análise faça sentido e para que os dados tenham um real significado. (BOELLSTORFF, 2013).

O thick data é moldado através da utilização de métodos de pesquisa qualitativa etnográfica que revelam as emoções das pessoas, histórias e modelos de seu mundo. (WANG, 2016). Tais métodos geram informações qualitativas que fornecem percepções sobre a vida emocional cotidiana dos consumidores, explicando questões de porquê. (REP, 2017).

Um exemplo prático dos resultados de análises que utilizam o thick data foi descrito por Rep (2017) ao descrever o case da Samsung. Após conduzir horas de entrevistas, análises de vídeos e de ouvir as pessoas, a Samsung conseguiu compreender que as pessoas viam a televisão muito mais próxima de um móvel do que um eletrônico e assim passaram a trabalhar o design dos seus produtos entendendo realmente o que eles representavam para os seus compradores. Outro exemplo é trazido por Mohammed e Fiaidhi (2019) que analisaram como a Lego, que estava muito próxima de um colapso nos início dos anos 2000 por não compreender o que os brinquedos representavam para seus clientes, conseguiu voltar a ser uma empresa de sucesso. Neste case, crianças de cinco grandes cidades foram estudadas para ajudar a Lego a entender melhor as necessidades emocionais das crianças em relação aos seus brinquedos. Depois de avaliar horas de gravações de vídeo destas crianças brincando com os legos, um padrão emergiu: as crianças eram apaixonadas pela experiência lúdica e pelo processo de brincar. Isso permitiu que a empresa destinasse seu foco para o local correto, afastando-se cada vez mais do fracasso.

Em comparação ao big data, o thick data difere pela abordagem qualitativa, obtendo dados etnográficos que permitem revelar contextos e emoções dos sujeitos analisados. O thick data depende do contexto social de conexões entre pontos de dados, enquanto o big data requer um processo de desenvolvimento geralmente executado por estatísticos e cientistas de dados, focando mais nos resultados do aprendizado de máquina. (MOHAMMED; FIAIDHI, 2019).

2.1.2 Big Data

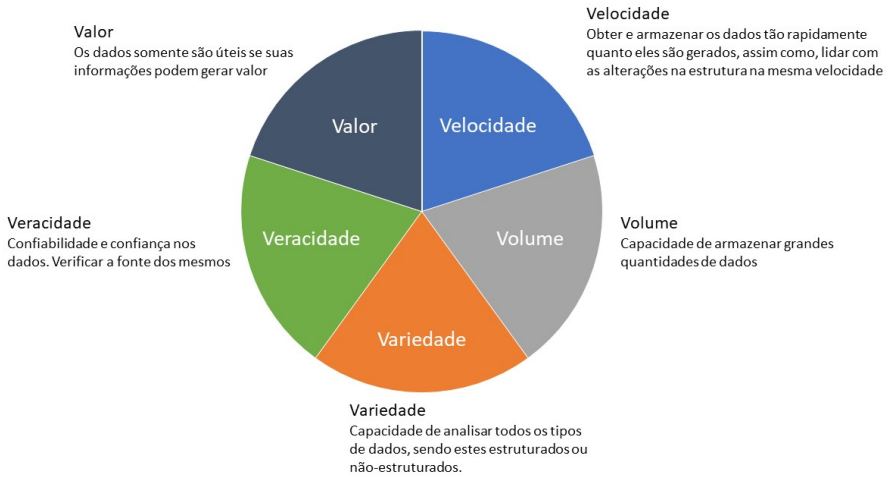
O big data está se tornando um novo foco de tecnologia na ciência e na indústria e motiva a mudança da tecnologia para modelos operacionais e arquitetura centrada em dados. (DEMCHENKO et al., 2014). O grande aumento na geração de dados globais faz com que o termo venha a ser utilizado principalmente para descrever a união desses dados em enormes conjuntos, mas o termo também está relacionado a quase todos os aspectos da atividade humana, desde o registro de eventos diários até pesquisas, design, produção de serviços digitais, entrega de produtos ao consumidor final entre outros. Além disso, o big data possibilita a descoberta de novos valores e auxilia na compreensão de novas análises. (CHEN; MAO; LIU, 2014).

Conforme citam Demchenko et al. (2014), há uma mudança de paradigma do host tradicional ou baseado em serviço para uma arquitetura centrada em dados e modelos operacionais com o big data. As propriedades utilizadas ao longo do trabalho são as definidas conforme Demchenko et al. (2014) e Khalid (2019) através dos 5Vs: volume, variedade, veracidade, valor e velocidade. Tais propriedades estão detalhas na Figura 1.

Destaca-se ainda que as tecnologias atuais, como cloud computing e conectividade de rede onipresente, fornecem uma plataforma para a automação de todos os processos de coleta, armazenamento, processamento e visualização de dados, fazendo com que o futuro do big data seja ainda mais próspero.

Wang (2016) destaca como o big data revela insights a partir de dados quantificados, através

Figura 1 – Propriedades do Big Data



Fonte: Adaptado de Demchenko et al. (2014) e Khalid (2019)

de técnicas que isolam variáveis para identificar padrões, mas que gera uma perda de resolução. Um exemplo prático do big data é descrito por Bornakke e Due (2018) como a onipresença de sensores que rastreiam os rastros digitais na internet com alta velocidade e de grande volume. No big data, os dados são o combustível que alimenta toda a arquitetura analítica. (DEMCHENKO et al., 2014).

2.2 Trabalhos Relacionados

Os trabalhos relacionados aqui descritos foram localizados através de uma revisão sistemática da literatura. Os detalhes dos procedimentos realizados estão descritos na seção 3, Materiais e Métodos. A Figura 2 ilustra o comparativo dos trabalhos relacionados, analisando o objetivo de cada artigo, quais fontes de dados foram utilizadas, se a proposta do autor mantinha o contexto dos dados utilizados, isto é, se o cenário de onde os dados foram obtidos poderia ser reconstruído para compreensão do significado deste dado no momento de sua coleta, se a proposta do autor estava preparada para combinar big data e thick data e, por último, qual o entregável de cada trabalho. Os detalhes de cada trabalho são apresentados a seguir.

Smets e Lievens (2018) tiveram como objetivo contribuir para o debate sobre a integração da etnografia e da ciência de dados, fornecendo uma ferramenta de pesquisa concreta para implantar essa integração. A ferramenta de pesquisa apresentada, Citizen Toolbox, combinou big data e thick data permitindo estudar as experiências dos ciclistas dentro de seu contexto. Ela não apenas relacionava big e thick data, mas também adicionava significado a cada tipo de dados individualmente, fornecendo um contexto mais amplo. Para tanto, a ferramenta de pesquisa possuía quatro estratégias de criação de sentido: dependência espaço-tempo, detecção de padrões em grandes conjuntos de dados, coleta de experiências subjetivas e combinação das

Figura 2 – Comparativo dos Trabalhos Relacionados

	Smets e Lievens (2018)	Mohammed e Fiaidhi (2019)	Fiaidhi (2020)	Presente Trabalho
Objetivo	Integração entre etnografia e ciência de dados e também uma ferramenta que disponibilizasse tal combinação a pesquisadores	Análise de comunidades em ambientes online para criar indicadores de saúde	Sistemas de saúde atuais não permitem visualizar os relacionamentos por completo dos dados dos pacientes	Demonstrar a possibilidade da combinação do thick data com o big data para atingir análises de dados mais acuradas e abrangentes
Fonte de Dados Utilizadas	Dados de uma cidade inteligente localizada na Antuérpia (Bélgica)	Redes sociais (geral)	Twitter	Literatura científica (cases teóricos), Twitter e Governo Federal (case prático)
Mantém o contexto dos dados originais	Sim	N/A	N/A	Sim
Combina big data e thick data	Sim	Sim	Sim	Sim
Entregável	Ferramenta para pesquisa denominada Citizen Toolbox	Modelo de sandbox com análises de thick data a partir de conversas do twitter	Framework para aprender as percepções dos pacientes nas redes de conversa do Twitter sobre saúde	Modelo de estrutura de dados que combina big data e thick data

Fonte: Elaborado pelo Autor

experiências com os dados objetivos para criar significado. A ferramenta foi compilada a partir de dois projetos, o primeiro é o Citizen Bike, que focou principalmente em como utilizar big data no processo de pesquisa. Já o segundo projeto, Be-Und, explorou o uso de dados contextuais para acionar interações pontuais com os participantes. O ponto forte dessa ferramenta era sua capacidade de combinar diferentes origens de dados e agir automaticamente sobre elas por meio do mecanismo de regras e do módulo de interação. A ferramenta permite que essas interações sejam acionadas dentro de um contexto específico, o que capacita o pesquisador na coleta de thick data.

Além disso, a caixa de ferramentas permite quatro tipos de construção de sentido: contextualização, semântica, análise e interpretação humana. Essas estratégias interagem entre si e, portanto, criam um valor adicional em termos de compreensão. Em outras palavras, essas estratégias permitiam aproximar o big data do thick data e também possibilitavam os pesquisadores a agir com base nos insights resultantes da integração de big e thick data. O Anexo A ilustra a arquitetura do Citizen Toolbox e os quatro tipos de construção de sentido citados anteriormente. O presente trabalho é similar ao de Smets e Lievens (2018) no que diz respeito a contribuição para o fomento do assunto thick data e a integração deste com o big data. A principal diferença é que durante o desenvolvimento deste trabalho, o objetivo foi construir uma arquitetura que além de combinar big data e thick data seja um modelo base, que é adaptável e pode ser utilizado em diversos casos, já Smets e Lievens (2018) forneceram uma ferramenta de pesquisa direcionada para as análises da cidade inteligente abordada no artigo, não gerando tamanha flexibilidade.

Já Mohammed e Fiaidhi (2019) desenvolveram uma abordagem que identificava conversas e aprendia o resultado da satisfação do consumidor a partir de tais discussões em mídias sociais. Inicialmente, identificavam os grupos de conversas e agrupavam em uma estrutura de grafo para analisá-las. Com base nesse grafo, era calculada uma variedade de indicadores de thick data, como a centralidade das conversas, a importância dos nós individuais do grafo e os influenciadores em cada uma dessas conversas. Através de algoritmos de detecção de comunidade, era possível dividir uma rede social em diferentes comunidades potencialmente sobrepostas. Em

uma tentativa de demonstrar a utilidade da estrutura de análise de thick data e os benefícios de usar esses dados qualitativos para os médicos aprenderem sobre as percepções dos pacientes e indicadores associados, como adesão a medicamentos ou reclamação sobre eventos adversos, desenvolveram uma sandbox para ser utilizada por médicos com o intuito de monitorar indicadores úteis das conversas com seus pacientes ou com outros usuários do Twitter sobre um determinado assunto. Em resumo, os autores apresentaram o paradigma crescente da análise de thick data e por fim destacaram a importância da identificação de comunidades de conversação nas mídias sociais.

O presente trabalho diferencia-se da abordagem dos autores principalmente por não restringir-se a aplicação específica em um local, como ocorreu com o Twitter utilizado no trabalho de Mohammed e Fiaidhi (2019). Além disso, Mohammed e Fiaidhi (2019) focaram em trazer os benefícios da utilização do thick data como uma nova fonte de análise, salientando o impacto que esta pode gerar, mas acabaram não abrangendo como o thick data poderia ser combinado a outra fontes.

Fiaidhi (2020) também voltou seu foco ao Twitter como uma fonte de insights, pois entendeu que era o ambiente mais dinâmico, com maior envolvimento e que permitia analisar a geolocalização do usuário. O Twitter possibilitou que as discussões dos pacientes se espalhassem e evoluíssem, adicionando um novo canal de percepções valiosas para a saúde. O artigo descreveu uma visão holística para identificar insights úteis do paciente no Twitter, relevantes para os pontos de dor do paciente, usando métodos baseados em insights qualitativos (IDLs). A abordagem era composta de cinco métodos IDLs, descritos a seguir, que forneceram aos médicos e profissionais de saúde insights significativos sobre as comunidades locais de seus pacientes e suas conversas. O primeiro método foi o desenvolvimento de um crawler de conversas que permitia a obtenção dos dados. O segundo método realizava a exploração das conversas da comunidade através da análise dos grafos desenvolvido no Neo4j. No método seguinte, ocorria a visualização das conversas em uma linha do tempo. Já o quarto método permitiu a análise de anomalias nas comunidades conversacionais. O quinto, e último método, concentrou-se em fornecer gráficos orientados ao contexto das comunidades de conversação, como aderência a gráficos de instruções médicas e conformidade de fornecimento de um serviço médico. Permitindo assim a identificação das redes de conversação através da API do Neo4j e a capacidade de inferir e visualizar as conversas de maior relevância.

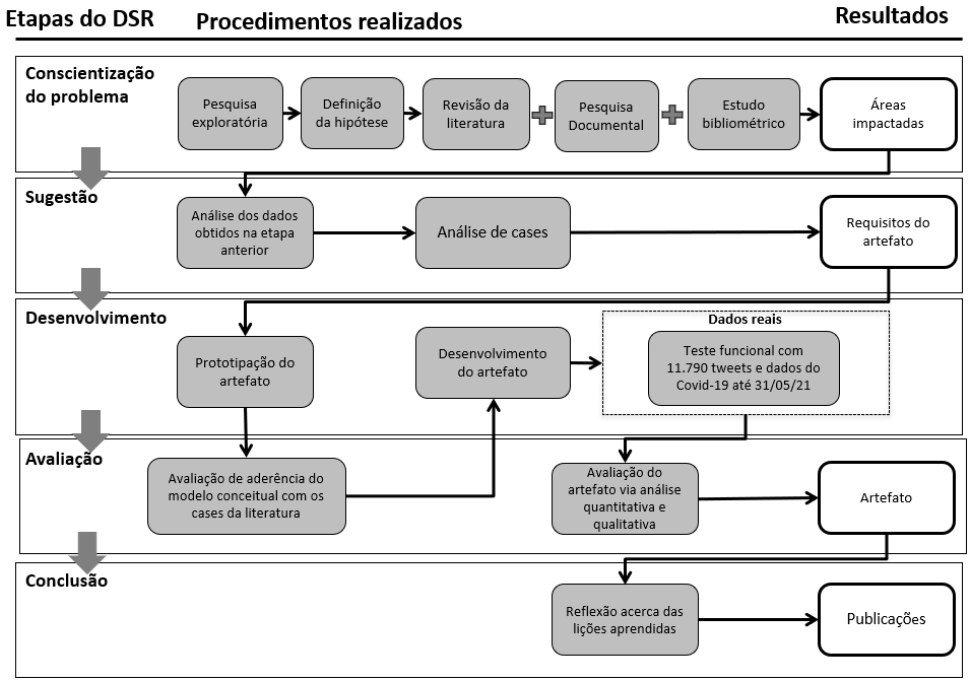
Os autores citaram dois métodos como trabalhos futuros que seriam a aplicação de machine learning para inferir o status de prognóstico dos casos de pacientes a partir das redes de conversação e ajustar o processo de modelagem de conversação capturado no método IDL de aprendizado de máquina para quando novos tweets chegassem com o uso do aprendizado por transferência. O Anexo B ilustra uma visão geral da estrutura desenvolvida. Novamente, a abordagem do presente trabalho expõe uma sugestão de modelo de dados ampla, que não se restringe a uma fonte específica como o Twitter e nem a uma área específica. Complementar a isso, Fiaidhi (2020) acabou não analisando como o thick data poderia ser combinado ao big

data, combinando os resultados encontrados em outras análises.

3 MATERIAIS E MÉTODOS

Para o desenvolvimento do trabalho foi aplicado o método Design Science Research (DSR). (KUECHLER; VAISHNAVI, 2008). O DSR possibilita resolver problemas de pesquisa de maneira mais eficaz e eficiente, além de ser capaz de trazer contribuições reais e práticas para o mundo. Sendo assim, foram seguidas as etapas propostas por Kuechler e Vaishnavi (2008), conforme apresentado na Figura 3 e detalhado a seguir.

Figura 3 – Etapas do DSR



Fonte: Elaborado pelo Autor

A primeira etapa do DSR consiste na conscientização do problema de pesquisa. Essa etapa foi realizada a partir dos seguintes procedimentos: pesquisa exploratória para buscar uma visão ampla do tema, definição da hipótese após entender as lacunas dos trabalhos já realizados, revisão da literatura juntamente com pesquisa documental para compreender o estado da arte no que se diz respeito à thick data e um estudo bibliométrico para analisar a relação quantitativa entre o número de publicações sobre big data frente ao thick data. A revisão sistemática da literatura e a bibliometria estão detalhadas nas subseções 3.1 e 3.2, respectivamente. Todos estes passos foram fundamentais para a conscientização do problema, encontrando diversas aplicações do modelo proposto. No final, a etapa gerou como resultado as principais áreas que podem ser impactadas positivamente com uma melhora no processo de análise dentro do big data, assim como, um extenso entendimento do tema.

A segunda etapa (sugestão) consiste na análise dos dados obtidos na etapa anterior, tais

como lacunas dos trabalhos relacionados, pontos de suma importância encontrados no estado da arte somado a uma análise de diversos cases, entre eles destacam-se Netflix e Lego. (MOHAMMED; FIAIDHI, 2019). Essa etapa resultou nos requisitos do artefato que tornaram-se posteriormente os design principles do modelo. É importante destacar que o design principle tem como principal característica estender os limites das capacidades humanas e organizacionais, criando assim artefatos inovadores. (HEVNER et al., 2004).

A terceira etapa (desenvolvimento) tem como primeiro objetivo a prototipação do artefato, isto é, aplicar os design principles no modelo. Após a etapa de avaliação do modelo conceitual, ocorre o desenvolvimento em si do artefato. Esse desenvolvimento já considera os pontos fortes e fracos encontrados na primeira avaliação, gerando os ajustes necessários. Com o modelo conceitual desenvolvido, ocorreu o teste funcional com dados reais obtidos através da API do Twitter (Twitter API, 2021) (qualitativos) e do Governo Federal (BRASILEIRO, 2021) (quantitativos).

A quarta etapa começa pela avaliação do modelo conceitual, analisando como o mesmo se comporta frente aos resultados esperados, isto é, se a estrutura desenvolvida é capaz de trabalhar com dados estruturados e não-estruturados, consegue manter os dados originais inalterados, além de conseguir analisar o contexto destes. Após, ocorre a avaliação final do artefato através de análises quantitativas e qualitativas, isto é, analisando o que pode ser visto de novo qualitativamente e como se comportaram os resultados numéricos da análise quantitativa. O resultado desta etapa é o próprio artefato, neste caso, o modelo que combina big data e thick data.

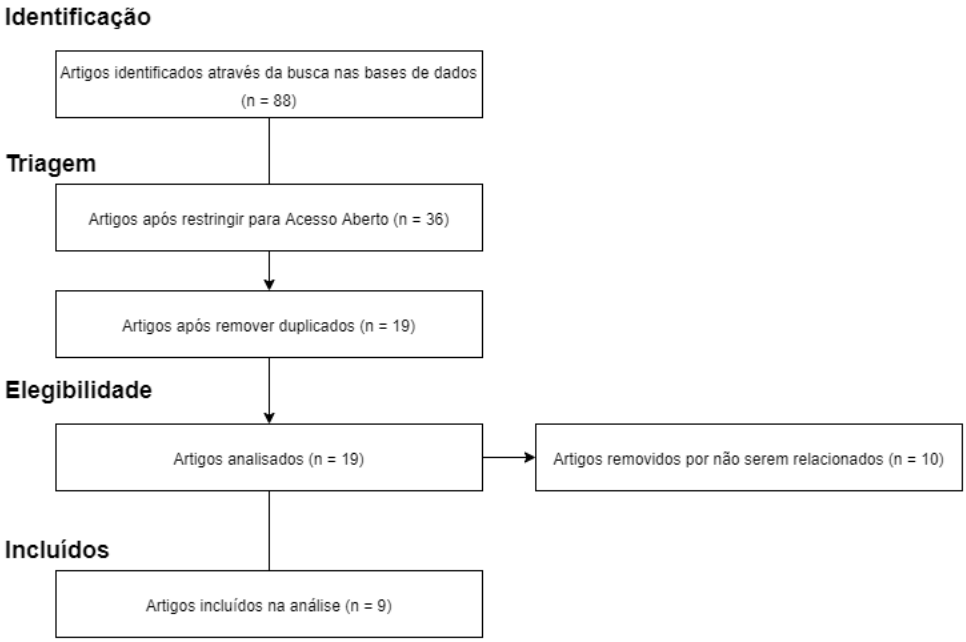
A quinta e última etapa é a conclusão. Aqui são realizadas as reflexões sobre as lições aprendidas e o resultado da etapa é a publicação do trabalho e do artefato desenvolvido.

3.1 Revisão Sistemática da Literatura

A revisão sistemática da literatura foi conduzida de acordo com o método PRISMA (MOHER et al., 2009), conforme ilustrado na Figura 4. A busca foi realizada nas bases de dados Scopus e Web Of Science. A string de busca aplicada inicialmente foi ‘thick data’ AND ‘big data’, contudo o número de resultados não era satisfatório: 6 na Scopus sendo todos de acesso aberto e 12 na Web of Science sendo apenas 7 de acesso aberto. Então, a string foi alterada apenas para ‘thick data’ e optado por analisar a relação entre o trabalho e o tema na fase de elegibilidade. Com a pesquisa somente por ‘thick data’, 43 resultados foram obtidos na base de dados Scopus e 45 na base de dados Web of Science. Então, aplicou-se o filtro por acesso aberto e foram obtidos 36 resultados. Destes 36 artigos, 17 foram removidos por serem duplicados. Os 19 trabalhos restantes foram analisados através do seu resumo para verificar a compatibilidade do tema e 10 foram removidos por não serem relacionados com o problema de pesquisa em questão. Os 9 artigos restantes foram lidos e foram escolhidos os 3 com maior proximidade ao tema para se utilizar como trabalhos relacionados. Para avaliar se um trabalho era relevante, os critérios avaliados foram a proximidade do tema do trabalho ou se o mesmo demonstrava al-

guma estrutura que combinasse thick data e big data que pudesse agregar ao estudo. O software Mendeley foi utilizado para gerenciar os artigos e para fazer a revisão dos mesmos.

Figura 4 – Fases da Revisão Sistemática de Acordo com o PRISMA



Fonte: Elaborado pelo Autor

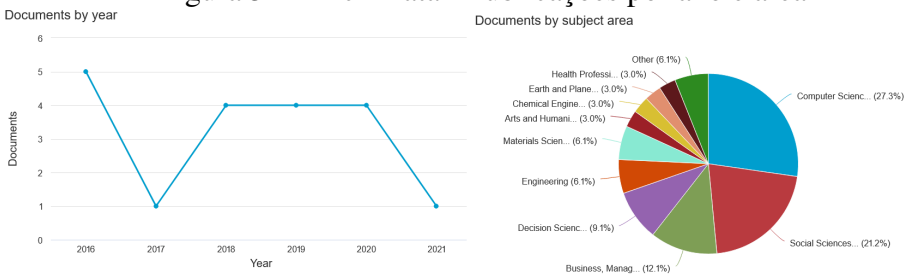
3.2 Bibliometria

Com o objetivo de avaliar e entender a relação dos avanços científicos nas áreas de thick data e big data, foi realizada uma bibliometria a partir da base de dados Scopus. A bibliometria é definida como um conjunto de métodos usados para estudar ou medir textos e informações, especialmente em grandes conjuntos de dados, além disso, diversas áreas de pesquisa utilizam métodos bibliométricos para explorar o impacto de seu campo ou a relação deste com os demais. (HENDERSON; SHURVILLE; FERNSTROM, 2009).

A pesquisa iniciou-se com os termos ‘thick data’, sem restrição alguma e o resultado foi 64 publicações. Após restringir o período de tempo entre 2016 e 2021, o número de resultados reduziu para 44. O último filtro aplicado foi referente ao acesso do material definido como acesso aberto culminando nos 19 resultados analisados pela quantidade de publicação por ano e por área, conforme ilustrado na Figura 5. Alterando a pesquisa para ‘big data’, os números são totalmente diferentes. São 105.619 resultados, que reduzem para 88.184 resultados ao filtrar o período de 2016 até 2021 e culminam em 23.556 publicações ao filtrar acesso aberto. Esses números estão ilustrados na Figura 6.

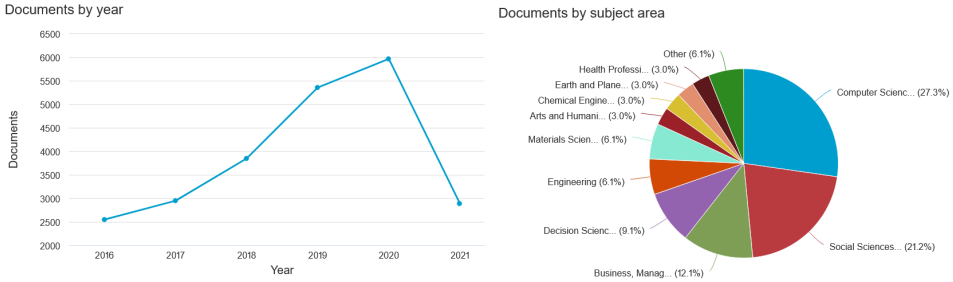
Os números indicam uma grande diferença quantitativa ao se comparar big data e thick data. Contudo, é importante citar que o termo thick data foi popularizado em 2016 por Wang (WANG, 2016) e ainda está em construção, podendo muitas vezes ter outras definições variando de autor

Figura 5 – Thick Data - Publicações por ano e área



Fonte: Base de Dados Scopus

Figura 6 – Big Data - Publicações por ano e área



Fonte: Base de Dados Scopus

para autor. De qualquer forma, a bibliometria reforça o tamanho do impacto que melhorias em big data podem gerar e também demonstra que ambos assuntos analisados estão relacionados com diversas áreas de aplicação.

4 CONSTRUÇÃO DO MODELO THICK E BIG DATA

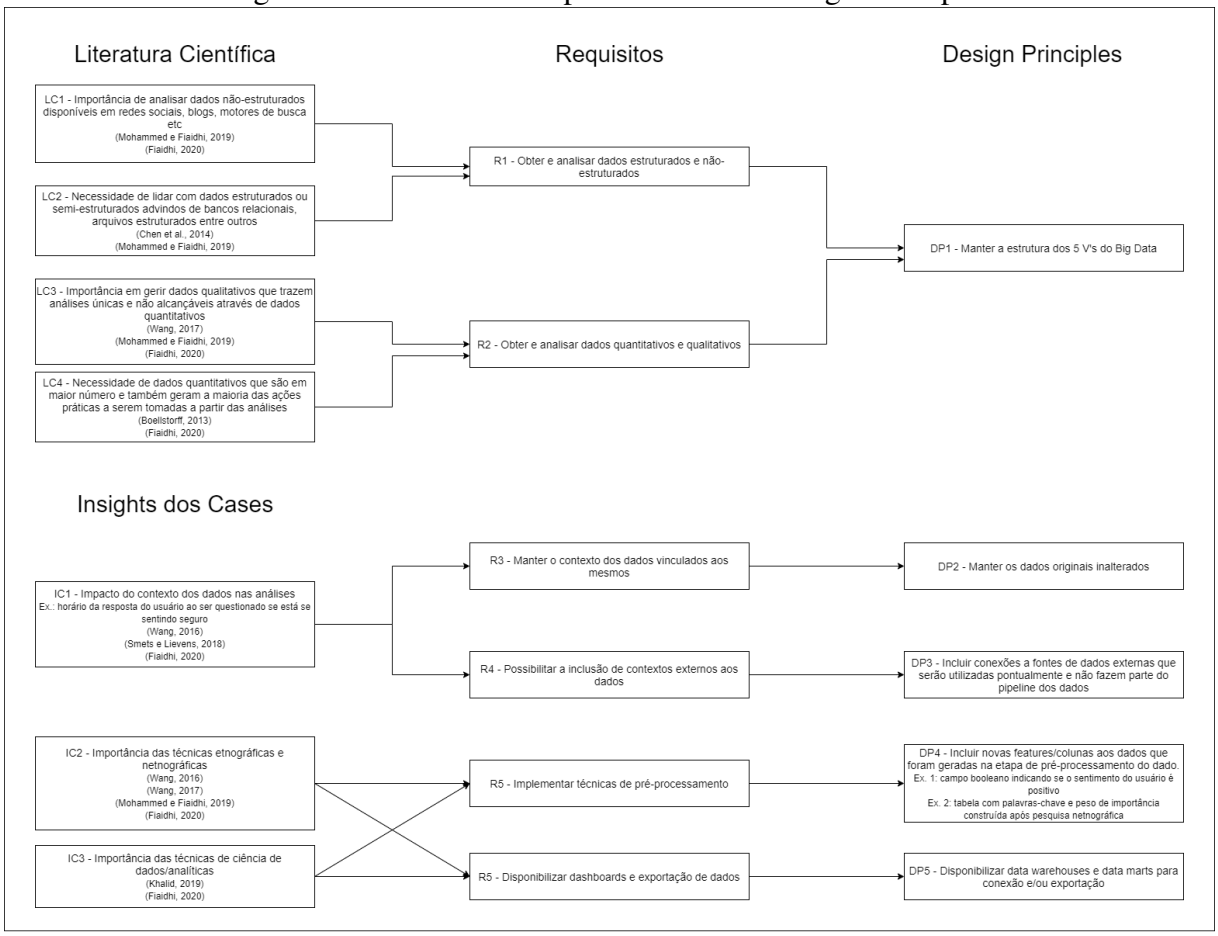
Esta seção está dividida em duas subseções, sendo a primeira direcionada ao detalhamento dos procedimentos utilizados para construção do artefato, desde a literatura, geração dos requisitos até os design principles. (HEVNER et al., 2004). E a segunda subseção detalha o modelo, conectando os design principles ao artefato desenvolvido.

4.1 Procedimentos para Derivar os Design Principles

A partir da literatura científica e dos cases analisados, foram obtidos como resultados da segunda etapa do DSR os requisitos do artefato. Os dois primeiros requisitos, R1 e R2, foram obtidos a partir do entendimento da necessidade de tratar os mais diversos tipos de dados, pois a estrutura deve estar preparada para conceber análises quantitativas e qualitativas afim de ter uma acurácia e abrangência maiores. O R3 e R4 abordam justamente a importância do contexto dos dados. Para se chegar nesses requisitos, a pesquisa documental foi de suma importância, pois trouxe diversos cases, somando-se aos cases encontrados na literatura científica que reforçaram

o impacto de analisar o dado dentro do contexto que este foi gerado. O R5 soma a importância da implementação de técnicas de ciência de dados como machine learning e deep learning a técnicas pouco utilizadas e que tem a grande capacidade de extrair informações de dados qualitativos, as técnicas etnográficas. (WANG, 2017). E o R6 materializa uma necessidade já conhecida de estruturas de big data atuais. (CENTER, 2018). Estes foram analisados e transformados nos design principles para a construção do modelo, a partir do entendimento da necessidade e também de análises de autores como Rietsche et al. (2018) e Meth, Mueller e Maedche (2015) que inspiraram o formato de apresentação do conteúdo. As etapas estão ilustradas na Figura 7 e detalhadas a seguir.

Figura 7 – Procedimentos para Derivar os Design Principles



Fonte: Elaborado pelo Autor

A parte esquerda da Figura 7 apresenta trechos da literatura e insights de cases que tendem a ser os principais responsáveis pelas taxas de falha e rentabilidade anteriormente. A partir destes, foram gerados requisitos que representam a materialização da necessidade posteriormente no modelo. Destes requisitos são derivados os design principles do modelo, isto é, cada item citado na última coluna deve ser encontrado no modelo proposto.

O design principle (DP) 1 indica a importância da atual estrutura de big data, ressaltando o quão sólida e consistente ela é. A literatura também mostrou como a arquitetura deve suportar

tantos dados estruturados quanto não-estruturados e também dados quantitativos e qualitativos, algo já conhecido no big data, destacado inclusive como o 'V' de variedade. (KHALID, 2019). Além disso, os demais Vs (velocidade, valor, veracidade e volume) são indiscutivelmente necessários em um modelo amplo e flexível, que busca combinar big data e thick data. (DEMCHENKO et al., 2014).

O DP 2 busca suprir lacunas de alguns modelos, onde após a execução do pipeline de transformação dos dados, não há uma maneira garantida de retornar aos dados originais. A necessidade em retornar aos dados obtidos originalmente ocorre quando uma análise só é vista posteriormente, isto é, após algum tempo de uso do modelo. Para exemplificar, pode-se utilizar um case abordado por Smets e Lievens (2018) onde é solicitado ao ciclista (usuário teste) se o mesmo está sentindo-se seguro ou não naquele momento. Caso a análise tivesse se baseado inicialmente nas respostas do usuário ou até em alguma outra combinação de dados, como a hora do dia, e o pipeline de dados mantivesse somente estas informações após sua execução, uma nova análise sobre o local que o ciclista estava não seria possível ou teria sido prejudicada pela perda de muitos dados. Além disso, esse princípio busca ajustar um comportamento comum do big data que é retirar o contexto dos dados e não poder trazer análises qualitativa e etnográficas pela falta deste contexto, conforme Pink et al. (2016) e Wang (2016).

O DP 3 está estreitamente ligado ao DP 2. Tomando como exemplo o mesmo case do ciclista, além de manter todos os dados originais dentro de seu contexto, pode ser necessária a combinação destes dados com outras fontes. Esta combinação pode ocorrer após o pipeline dos dados, pois será utilizada pontualmente e não há necessidade de guardá-los. Naquele exemplo, além de querer analisar a localização do usuário, estes dados podem ser cruzados com dados meteorológicos e então um novo padrão pode surgir na análise: uma região que se considerava insegura pelo seu alto índice de respostas indicado pode ser reavaliada quanto a ocorrência ou não de chuvas nos locais e os horários da solicitação feita ao usuário.

O DP 4 surge do entendimento da importância da aplicação tanto de técnicas de ciência de dados quanto de técnicas de etnografia. (WANG, 2016). O design principle em questão está relacionado a aplicação destas técnicas e do armazenamento do resultado delas como uma nova coluna da tabela ou uma nova feature do conjunto em questão. Um exemplo prático seria a inclusão de uma coluna 'Cliente Muito Importante' numa tabela relacionada a clientes e suas compras, onde o valor booleano que define a importância do cliente foi obtido através de técnicas etnográficas que analisaram o comportamento do cliente durante o ato da compra. Outro exemplo seria agrupar esse cliente através de técnicas de machine learning voltadas para criação de clusters e indicar a qual grupo ele pertence através de uma coluna nova nos dados.

Já o DP 5 é uma característica também já conhecida de estrutura de big data, a camada de disponibilização dos dados para ferramentas analíticas. Os data warehouses (DW) contém todas as informações para tomada de decisão e são o armazenamento central dos dados. Enquanto isso, os data marts são construídos a partir dos DWs e são focados em uma aplicação ou área de negócio específica. (ALAOUI et al., 2019). Este design principle aborda também a

possibilidade da exportação dos resultados.

Através destes DPs, diversas lacunas dos trabalhos relacionados e dos cases analisados são preenchidas. Eles guiam o modelo para uma estrutura flexível e capaz de ser aplicada as mais diversas áreas. Além disso, eles conseguem extrair o contexto dos dados qualitativos e gerar assim análises mais abrangentes e, por vezes, também mais assertivas. A geração destes DPs e a posterior inclusão deles no modelo, possibilita a utilização deste modelo em diversas áreas, pois o mesmo se torna flexível, servindo como base para estruturação dos dados. Essa contribuição é importante e não foi encontrado algo semelhante nos trabalhos relacionados.

4.2 Modelo Proposto

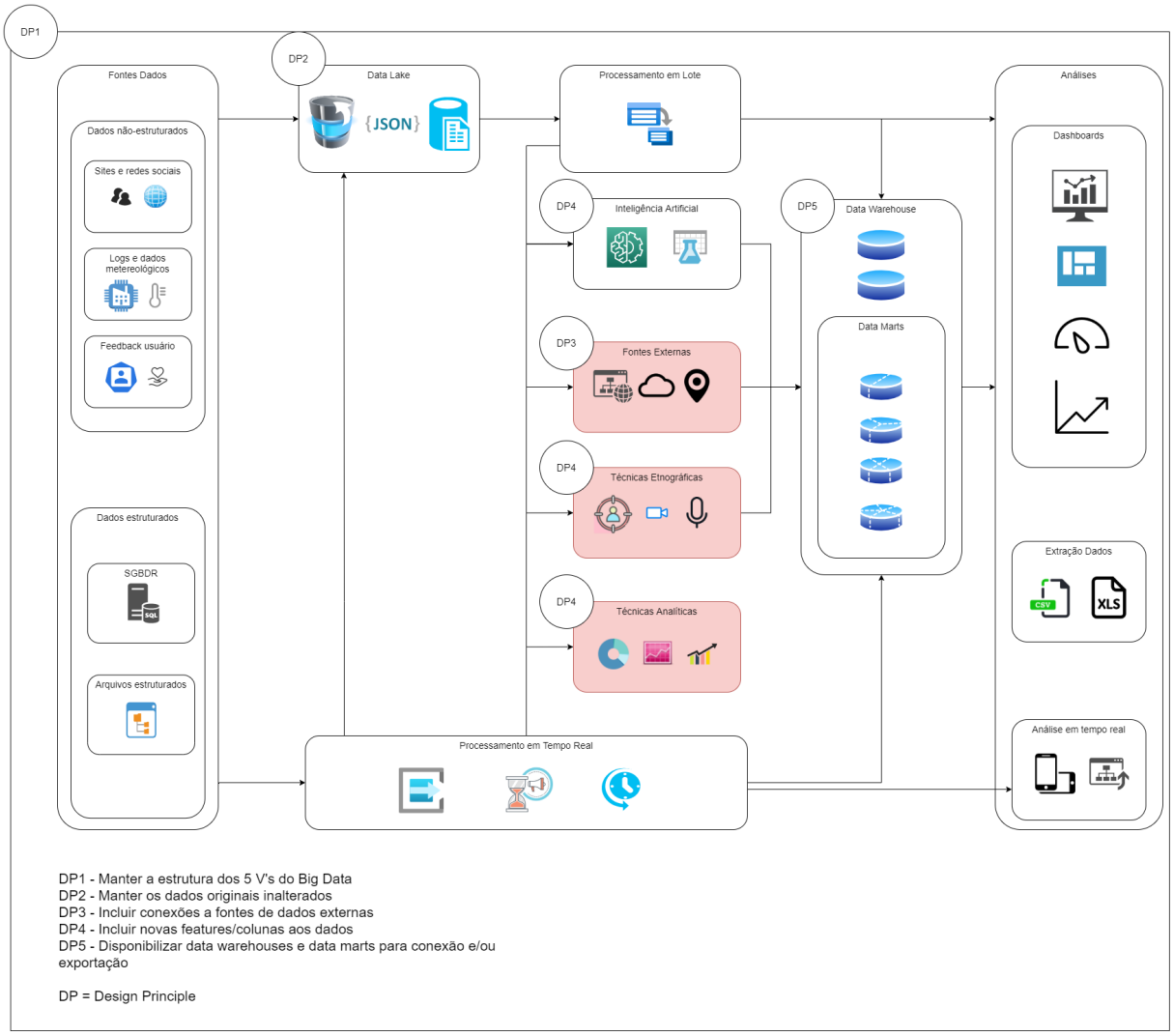
O modelo proposto é baseado em uma arquitetura de big data já consolidada tanto academicamente (DEMCHENKO et al., 2014) quanto no mercado. (CENTER, 2018). A partir dessa arquitetura e dos design principles, foram adicionados novos recursos que buscam suportar e potencializar a análise de dados qualitativos. Estas novas etapas estão destacas em vermelho no modelo, que está ilustrado na Figura 8.

A primeira estrutura apresentada, à esquerda na Figura 8, são as 'Fontes de Dados' onde dados não-estruturados e dados estruturados estão inseridos. Os dados-não estruturados podem ser obtidos através de sites, redes sociais, logs, dispositivos inteligentes, dados meteorológicos etc. Este tipo de dado comumente é o que apresenta mais características e possibilidades de análises qualitativas. Tais características são necessárias para aplicação de técnicas etnográficas, netnográficas ou técnicas qualitativas no geral. Abaixo dos dados não-estruturados apresenta-se os dados estruturados, a fonte mais utilizada para análises pelo big data. (WANG, 2016). A existência dessas duas fontes são de suma importância para a potencialização das análises, pois é necessário ter uma visão completa da situação a ser analisada.

'Data Lake' é a solução proposta para o DP 2, pois ele torna-se um repositório central de dados capaz de armazenar grandes quantidades de dados, podendo estes ser estruturados, semi-estruturados ou completamente não-estruturados. (SINGH; AHMAD, 2019). A ideia é armazenar todos os dados exatamente como foram obtidos em suas fontes. Com isso, são resolvidos os problemas encontrados na literatura e também nos cases analisados referente a limitações de análises e, principalmente, a capacidade de manter o contexto do dado atrelado a este. O 'Processamento em Lote' refere-se ao armazenamento dos dados disponíveis no data lake em data warehouses e, posteriormente, em data marts. Nesta etapa são realizados os tratamentos necessários, assim como, denormalizações para construir uma estrutura focada na leitura de dados por ferramentas analíticas. Esse processamento pode ou não contar com a aplicação de técnicas de pré-processamento ou fontes externas.

Paralela a essa estrutura, existe a capacidade de 'Processamento em Tempo Real'. Como pode ser analisado na Figura 8, os dados extraídos podem tanto ser apenas armazenados em um data lake, quanto serem analisados em tempo real, dependendo da necessidade. A ideia

Figura 8 – Modelo Proposto



Fonte: Elaborado pelo Autor

é que os dados analisados em tempo real também sejam inseridos no data lake para posterior composição do data warehouse.

A estrutura referente à 'Inteligência Artificial' (IA) está relacionada com o DP 4, isto é, armazenar pré-processamentos feitos com técnicas de IA, tais como técnicas de machine learning ou deep learning. O armazenamento desses valores, quando possível, reduz muito o esforço das ferramentas de análise que não precisarão aplicar técnicas internamente, otimizando assim o processamento. Em contrapartida, isso também requer um armazenamento maior, gerando um trade-off que deve ser avaliado.

Analisa-se agora as principais contribuições do presente trabalho, são elas: Fontes Externas, Técnicas Etnográficas e Técnicas Analíticas. Todas já foram detalhadas e exemplificadas na subseção anterior, contudo, cabe ressaltar que este é o ponto chave para o processo de combinar o thick data ao big data. É de suma importância o entendimento de que para ocorrer a extração máxima de informações dos dados qualitativos, estes precisam ser compreendidos. Tal

compreensão ocorre através de uma pré-análise, comumente realizada por uma pessoa com domínio do assunto. A falta dessa análise é um dos fatores que levam a falhas em projetos de big data, pois atualmente espera-se que os algoritmos de IA consigam extrair todas informações necessárias e nem sempre isso ocorre. Sendo assim, o papel do analista é fundamental na maioria dos casos.

A caixa 'Fontes Externas' refere-se ao DP 3, permitindo maior flexibilidade para o cruzamento dos dados já obtidos, e tratados, com fontes externas que possam complementar ou construir o contexto necessário. As 'Técnicas Etnográficas' são a primeira parte da solução do DP 4. Mais do que apenas ilustrar a capacidade de uma etapa de pré-processamento que se conecte ao data warehouse e seus data marts, essa etapa é a mais importante quando se trata em utilizar thick data. As técnicas etnográficas tem a capacidade de extrair diversas análises a partir de dados qualitativos e, portanto, são a chave para o funcionamento do modelo proposto. As 'Técnicas Analíticas' complementam a solução do DP 4. Embora possam ser muito próximas as técnicas de IA, elas estão especificamente declaradas por serem obtidas comumente a partir de análises exploratórias e entendimento dos dados, realizados, neste caso, por uma pessoa e não apenas com algoritmos de inteligência artificial. É nessa etapa que o cientista de dados, ou mesmo um especialista no assunto, pode explorar os dados, combinar diversas fontes externas, encontrar alguns padrões e então criar um pipeline que armazena estes novos dados.

O DP 5 está ilustrado através dos data warehouses e dos data marts. Essa é uma estrutura já consolidada (ALAOUI et al., 2019) e está inserida para complementar o modelo conceitual, assim como, as 'Análises'. Este DP ressalta as capacidades de conexão de ferramentas analíticas, principalmente ferramentas de business intelligence, de extração das bases para um novo pipeline e de ferramentas que consigam analisar em tempo real os dados que assim foram tratados.

5 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Para analisar a aderência do modelo proposto, foram utilizados dois cases obtidos durante a revisão sistemática de literatura e desenvolvido um case prático para observar a aplicação do modelo com dados reais.

5.1 Cases Observados na Literatura

O primeiro case foi obtido a partir do trabalho de Mohammed e Fiaidhi (2019). Os autores citam como a Netflix combinou as grandes quantidades de dados que possuíam dos hábitos de visualização dos clientes (big data) com estudos etnográficos para entender melhor o comportamento de cada assinante. Nesse case, já é identificada necessidade de possuir uma estrutura de big data que suporte trabalhar dados quantitativos, isto é, o DP 1 do modelo. Estes dados quantitativos por si só conseguiam responder questões do tipo: quem está assistindo? O quê?

Quando? Mas haviam outras perguntas que eles queriam responder: por que o assinante estava assistindo? Quais necessidades ele tinha que a Netflix supria? E como essas necessidades eram entregues? Essas perguntas foram respondidas através de estudos etnográficos, a principal proposta do modelo obtida através do DP 4. O impacto de utilizar thick data para complementar o big data fez com que a gigante do streaming entendesse que o seu público tem preferência por assistir diversos episódios em sequência, ou seja, 'maratonar' as séries. Com esse entendimento, a empresa conseguiu direcionar corretamente sua criação de conteúdo, compras de direitos de reprodução e até a maneira com que disponibiliza o conteúdo.

Além dos DPs detalhados acima, os DP 2 e 5 não foram claramente citados, mas são comumente encontrados em estruturas de big data atuais. (CENTER, 2018). E o DP 3 não está relacionado a este case em específicos, pois esta análise não utilizava fontes externas.

O segundo case foi extraído do trabalho de Smets e Lievens (2018) onde os autores buscaram avaliar como a experiência do ciclista poderia ser melhorada. Eles obtiveram os dados de geolocalização do usuário a partir do dispositivo que criaram e, complementar a estas informações, armazenavam dados de sensores que estavam instalados em vários pontos da cidade conseguindo assim capturar o contexto dos dados. Neste ponto, já se mostrou a necessidade de suportar grandes quantidades de dados através de uma estrutura de big data e também com alta velocidade de gravação (DP 1), pois para uma análise assertiva esses dados deveriam ser obtidos com pouca diferença de tempo e em grandes quantidades. O entendimento da importância do contexto dos dados, que gerou a necessidade de instalar sensores é o que deve ser destacado. O DP 4 do modelo proposto trabalha justamente em analisar o contexto dos dados, logo, ter esse contexto já armazenado faz o que todos os dados necessários para análise completa estejam corretamente armazenados. A partir destes dados, eles chegaram a alguns insights gerais que foram aprofundados através de uma pesquisa qualitativa adicional, gerando assim um profundo entendimento da experiência dos ciclistas. Novamente, destaca-se a importância de técnicas etnográficas, que nem sempre serão aplicadas somente ao dados existentes, mas podem ser aplicadas em cima dos novos horizontes que os dados geraram.

Neste case, o DP 2 e DP 5 também não são citados por serem parte da arquitetura não explorada pelos autores, contudo são de suma importância para uma estrutura completa de análise. Além disso, o DP 3 foi detalhado na subseção anterior adicionando uma nova análise justamente neste case, onde o input de um novo dado, como previsão do tempo, poderia gerar um novo entendimento dos possíveis padrões encontrados.

5.2 Case Prático

O case prático busca demonstrar a importância de combinar dados qualitativos e quantitativos. Complementar a isso, o case também reforça os design principles do modelo e a aplicação do mesmo numa estrutura de análise de dados. Foram levantados alguns possíveis temas para o case e, devido ao momento que vive-se definiu-se que a análise seria referente ao Covid-19.

Frente a este tema, foram levantadas algumas propostas que buscavam analisar qualitativamente e quantitativamente eventos específicos como o impacto das alterações de bandeiras de determinada região ou o impacto de datas comemorativas ocorridas no final de 2020. Contudo, uma das principais limitações do case prático é relacionado a dimensão temporal dos dados qualitativos impedindo que tais análises fossem realizadas. Com isso, novos cases foram pensados e dentre estes definiu-se o objetivo de analisar a influência das manifestações ocorridas no dia 1 de Maio de 2021 (Dia do Trabalhador) pró-Bolsonaro. (G1, 2021). Para a realização deste case, foram feitas análises qualitativas e quantitativas, detalhadas a seguir.

5.2.1 Análise Qualitativa

A análise qualitativa tem como objetivo explorar os dados, buscando compreender qual é o objetivo que cada discurso buscava trazer ao ser dito. Para isso, foi feita uma análise amostral dos tweets seguido da aplicação de uma técnica de análise de conteúdo adaptada de Bardin e Reto (2016). O Twitter foi escolhido para ser a fonte dos dados qualitativos, pois ele é uma ótima opção para ser a fonte primária das informações e também é imensamente rico para análises. (MARWICK, 2014). Contudo, é necessário citar a principal limitação do mesmo referente ao período de busca. Devido ao tipo de usuário especificado pelo Twitter que se encaixa no presente trabalho, a busca dos dados era limitada ao período de 7 dias anteriores ao dia da chamada da API, conforme ilustrado no Apêndice A . (Twitter API, 2021).

Para obtenção e armazenamento dos tweets, foi desenvolvido um algoritmo em Python que realiza a coleta dos tweets através da API oficial (Twitter API, 2021) no período de 29/04/21 à 03/05/21 e armazena os mesmos no MongoDB. (MongoDB, 2021). Foi escolhido o MongoDB devido sua versão gratuita de instalação e armazenamento em cloud, além de possuir vasta documentação disponível. A orquestração das chamadas no método main para os métodos e o fluxo de obtenção estão ilustrados pela Figura 9. Durante a obtenção dos dados, outra limitação se fez presente. Só era possível obter 100 resultados por solicitação na API, conforme ilustrado no Apêndice B. Esta característica estava descrita na documentação, mas não havia justificativa para tal. Com isso, caso fossem enviadas diversas solicitações passando um período de um dia, por exemplo, o mesmo tweet poderia ser gravado repetidamente na base de dados ou caso fosse implementada essa validação, o número de tweets coletados ficaria abaixo do esperado. Para solucionar isso, foi definido que 100 tweets por hora seria o suficiente para uma análise assertiva e criado um array com o dia e a hora de solicitação, conforme Apêndice C.

Como o case buscava entender o impacto das manifestações, foi definido que seria utilizado um período de cinco dias, sendo dois antes do evento, o dia do evento e dois dias posteriores ao evento. A partir desse objetivo, também se criou a premissa para a query de pesquisa que é retornar todos tweets com a palavra ‘covid’. Devido ao contexto, foi especificada a linguagem português e também se optou por desconsiderar retweets, pois esses poderiam se referenciar a um mesmo assunto, impactando assim na análise do cenário geral. O número máximo de

resultados, assim como a query construída para a pesquisa, as datas e a construção da URL estão ilustradas na Figura 10. Ao final da coleta, a base de dados contava com 11.790 tweets, sendo 4.744 dos dias que antecediam o evento, 2.381 do dia do evento e 4.665 dos dois dias após o evento. A escolha de dois dias antes e dois dias após foi feita para conseguir ver o impacto no contexto social mais amplo e não só no dia antecessor ou posterior ao evento. A diferença do número de tweets dos dias que antecedem e sucedem o evento frente ao número de tweets do dia do evento não causam impacto, pois a análise se baseia no rankeamento da ocorrência e não no número de ocorrências. É importante destacar que os trabalhos relacionados não informaram se armazenaram os tweets e como teriam armazenados, apenas Marwick (2014) cita ferramentas prontas que utilizou para seu trabalho, não tendo desenvolvido estas.

Figura 9 – Método Main do Algoritmo para Obtenção e Armazenamento dos Dados

```
if __name__ == "__main__":
    inicio = datetime.now()
    print("Data de início: " + str(datetime.now()))


    headers = create_headers(bearer_token)
    db = connect_mongoDB()
    collection = db.covid_impacto_manifestacao

    i = 0
    while(i<len(times_to_search)-1):
        url = create_url(times_to_search[i], times_to_search[i+1])
        print("Consultando: " + str(times_to_search[i]) + " - " + str(times_to_search[i+1]) + " às " + str(datetime.now()))
        json_response = connect_to_endpoint(url, headers)
        try:
            tweets = json_response['data']
            insert_many_mongoDB(collection, tweets)
        except:
            print("Sem dados para o período: " + str(times_to_search[i]) + " - " + str(times_to_search[i+1]))
        finally:
            i = i + 1

    fim = datetime.now()
    print("Data de fim: " + str(datetime.now()))

    print("Tempo de execução: " + str(fim-inicio))
```

Fonte: Elaborado pelo Autor

Figura 10 – Criação da URL para chamada da API do Twitter

```
def create_url(start_time_to_search, end_time_to_search):
    query = "covid lang:pt -is:retweet"

    tweet_fields = "tweet.fields=author_id,geo,created_at"
    start_time = "&start_time=" + start_time_to_search
    end_time = "&end_time=" + end_time_to_search
    max_results = "max_results=100"
    url = "https://api.twitter.com/2/tweets/search/recent?query={}&{}&{}".format(query, tweet_fields, max_results) + start_time + end_time
    return url
```

Fonte: Elaborado pelo Autor

Com os dados armazenados, foi desenvolvido outro algoritmo em Python que fazia a leitura do MongoDB, transformava os tweets em um único texto e então fazia a tokenização deste. A tokenização é o princípio de isolar as palavras de uma sentença através pontos de isolamento. (KIBBLE, 2013). Além da tokenização, nessa etapa eram retiradas as stop words, pontuações e palavras que não faziam sentido para análise, como palavras que compõe URL e abreviações de stop words, conforme sugerido na própria documentação da biblioteca NLTK. (NLTK, 2021). A construção da lista está representada no Apêndice D. Com os tokens já construídos, é realizada

a contagem da ocorrência de cada palavra e gerado um arquivo de texto com as 100 palavras de maior ocorrência, seu ranking e a quantidade de usos, conforme ilustrado na Figura 11. Também são geradas word clouds desse resultado para se obter uma análise visual, conforme Apêndices E até J. Compreendendo algumas limitações da análise da palavra como um todo, foi realizado o processo de stemming nos tokens e refeita a contagem e geração de arquivos. (KIBBLE, 2013). Sendo assim, a análise qualitativa consegue gerar as cem palavras mais utilizadas dentre os tweets coletados e também os cem stemming mais utilizados, tanto estatisticamente como visualmente, tornando-se uma adaptação da análise de conteúdo onde é feita uma leitura e análise do texto e então relacionado este a uma legenda. (BARDIN; RETO, 2016). Esses dados foram analisados combinando-se a análise quantitativa.

Figura 11 – Métodos de Contagem de Ocorrência e Ordenação destas Ocorrência no Arquivo de Saída

```
def dict_frequencies(wordlist_without_sw):
    wordfreq = []
    dict_frequencies = {}
    for w in wordlist_without_sw:
        wordfreq.append(wordlist_without_sw.count(w))
        dict_frequencies[w] = wordlist_without_sw.count(w)
    return dict_frequencies

def most_frequency(dict_freq, arq, top_n):
    d = Counter(dict_freq)
    d.most_common()
    time = datetime.now()
    f = open("TopWords/covid_impacto_manifestacao_durante_" + arq + " - " + str(time.year) + "-" + str(time.month) + "-" + str(time.day) + "_"
    + str(time.hour) + "-" + str(time.minute) + "-" + str(time.second) + ".txt", "a")
    posicao = 1
    for k, v in d.most_common(top_n):
        f.write('%i - %s: %i \n' % (posicao, k, v))
        posicao += 1
    f.close()
```

Fonte: Elaborado pelo Autor

5.2.2 Análise Quantitativa

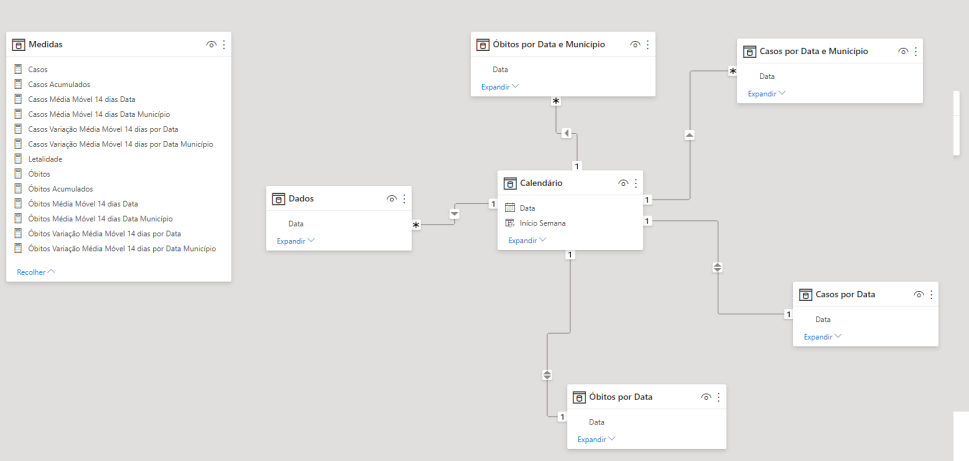
A análise quantitativa tem como objetivo analisar os possíveis impactos levantados na etapa de análise qualitativa. Para isso, foi utilizada uma ferramenta analítica específica que tem capacidade para analisar os dados obtidos e identificar padrões caso estes existam. A ferramenta escolhida foi o Power BI, pois é um software que possui uma versão free (Power BI Desktop) e tem capacidade técnica para obter e analisar os dados obtidos. (BI, 2021).

A base de dados quantitativa é um arquivo de extensão csv disponibilizado pelo Governo Federal. (BRASILEIRO, 2021). A mesma foi exportada contendo os dados até o dia 01/06/2021. Dentre as informações presentes na base, foram utilizadas: a data de referência, o número de casos novos, o número de óbitos e o município. Ao analisar a base de dados, a primeira transformação se fez necessária ao se observar que o mesmo dado estava incluído na região 'Brasil' e na sua região de ocorrência, sendo que a região 'Brasil' era uma agregação das demais áreas. Portanto, foram filtrados os dados que não pertenciam a região 'Brasil'. Além disso, também se optou por remover os dados pertences ao dia 01/06, pois a análise não utilizaria estes dados e a visão mensal dos casos ficaria divergente.

Com os dados tratados, foram desenvolvidas as medidas necessárias para análise e detalha-

mento dos pontos levantados na análise qualitativa. Medidas é a nomenclatura utilizada para definir as fórmulas e cálculos dentro da ferramenta. Destaca-se nessa etapa a necessidade de acompanhamento dos casos, casos acumulados, óbitos, óbitos acumulados, média móvel de casos, variação da média móvel e a análise dessas medidas pelas dimensões de município e data. Devido a granularidade da tabela disponibilizada, para análise por data e município, a geração de tabelas intermediárias dentro da ferramenta foi necessária, fazendo assim com que a média fizesse sentido frente ao dado em questão. O modelo de dados, assim como as medidas criadas, estão ilustrados na Figura 12.

Figura 12 – Modelo de Dados e Medidas criadas no Power BI



Fonte: Elaborado pelo Autor

5.2.3 Relação do Modelo Proposto com Case Prático

O modelo proposto foi utilizado com base para a realização do case prático. A primeira relação que pode ser encontrada é relacionada ao DP1 e é referente as fontes de dados utilizadas. Durante a realização do case, conforme já detalhado, foram utilizados dados qualitativos obtidos a partir da API do Twitter (Twitter API, 2021) que encontravam-se na forma semi-estruturada no formato JSON e dados quantitativos disponibilizados pelo Governo Federal (BRASILEIRO, 2021) que encontravam-se em extensão .csv.

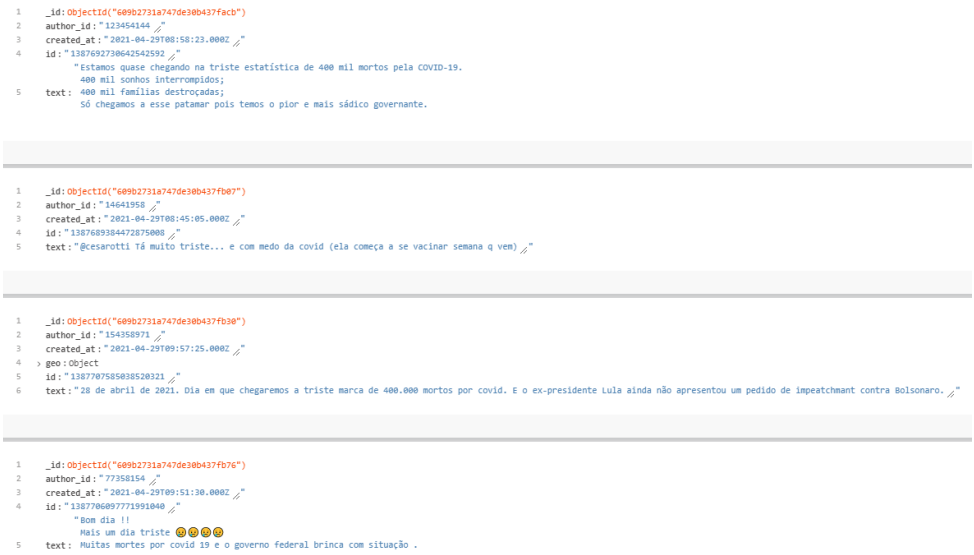
O armazenamento destes dados exatamente como foram obtidos relaciona-se com o DP2. Embora não tenha sido necessária a construção de um data lake, os mesmos foram armazenados sem alteração alguma para garantia da consistência do contexto destes e também para possíveis análises não identificadas no primeiro momento. O DP 4 esteve presente tanto nas análises qualitativas realizadas no algoritmo em Python como nas análises desenvolvidas dentro do Power BI.

Para esse case, não houve a necessidade de busca de dados em fontes externas (DP 3) ou de geração de data warehouses ou data lakes (DP5), mas estes podem ser aplicados numa futura evolução do case.

5.2.4 Análise e Discussão do Resultados do Case Prático

A análise qualitativa iniciou-se com a leitura de uma amostragem aleatória de tweets por período referente ao evento. Nessa primeira iteração, foi possível identificar os principais assuntos que permeavam entre o sentimento de tristeza referente as perdas decorrentes da Covid-19, a esperança e aguardo pela vacina e o posicionamento político de cada autor, tendo como destaque críticas referentes à atual presidência. Um exemplo dessa análise está ilustrado na Figura 13.

Figura 13 – Exemplos de Tweets Obtidos por Amostragem



Fonte: Elaborado pelo Autor

A análise dos arquivos de textos gerados pelo código Python reforça a impressão obtida na análise amostral, principalmente referente a vacinação e as menções ao Presidente Jair Bolsonaro. A palavra ‘vacina’ fica entre quarto e sexto lugar durante todo o período analisado, assim como, o stemming ‘vacin’ que permanece em segundo lugar de palavra mais dita. A palavra ‘bolsonaro’ também apresenta constância, variando entre décimo e décimo quarto. A Figura 14 ilustra as palavras e stemmings citadas nesta análise.

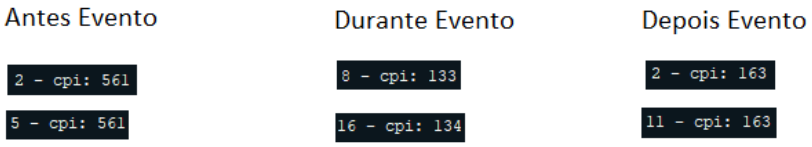
A primeira situação destacada pela análise de ocorrência de palavras e stemming é referente a CPI. Nos dias que antecedem as manifestações, a palavra estava posicionada em segundo lugar e o stemming em quinto lugar e caem, respectivamente, para oitavo e décimo sexto durante o dia 1º de maio. Depois das manifestações, a palavra retorna para a segunda colocação e o stemming fica em décimo primeiro. Com essas informações, já é possível perceber que um dos impactos dos eventos foi a alteração do foco dos comentários dos autores dos tweets, conforme ilustrado na Figura 15.

A segunda análise é referente a sensibilização e espanto frente ao número de casos da covid. O stemming ‘cas’, referente em sua grande maioria as palavras ‘caso’ e ‘casos’, vai de décima

Figura 14 – Análise de Palavras e Stemming em todo Período de Análise

Fonte: Elaborado pelo Autor

Figura 15 – Análise da Palavra e do Stemming ‘cpi’ ao longo do Evento



Fonte: Elaborado pelo Autor

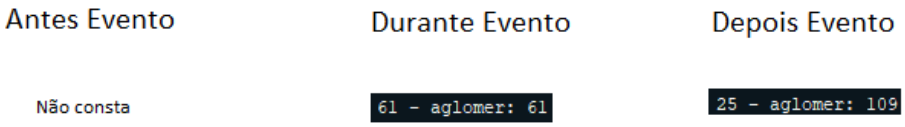
primeira colocada antes das manifestações para quarta e quinta, durante a após respectivamente como ilustrado na Figura 16.

Figura 16 – Análise do Stemming ‘cas’ ao longo do Evento

Fonte: Elaborado pelo Autor

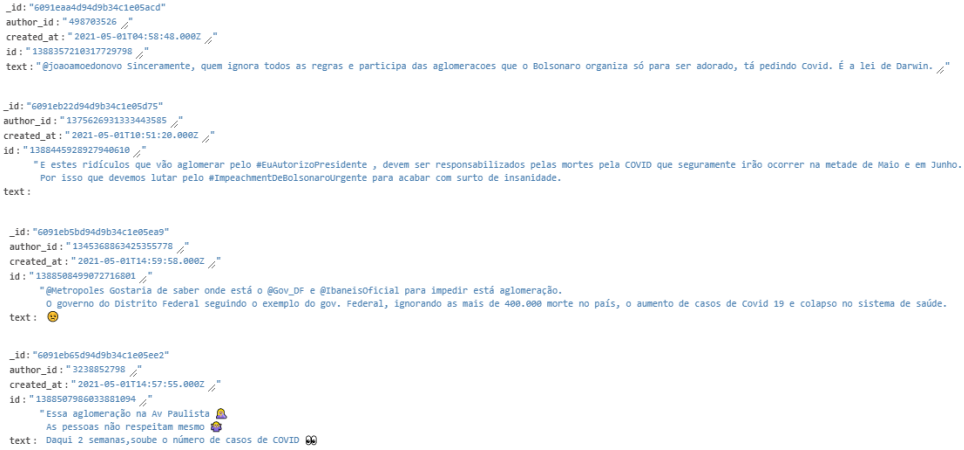
A principal análise é referente as críticas sobre aglomerações. Embora seja com certa frequência que as pessoas comentam sobre o assunto, nem a palavra nem seu stemming estava rankeados nos dados referentes aos dias que antecedem as manifestações, 29 e 30/04/2021. Contudo, 'aglomer' surge nos dados coletados no dia 01/05/2021 na posição 61º e depois para 25º nos dias após o evento, conforme ilustrado na Figura 17. Ao se analisar amostras de tweets referente a esse assunto, o que mais se comentava era a desconsideração das pessoas que estavam participando das manifestações e, principalmente, a grande possibilidade de um aumento no número de casos nos dias seguintes, ilustrado na Figura 18.

Figura 17 – Análise do Stemming 'aglomer' ao longo do Evento



Fonte: Elaborado pelo Autor

Figura 18 – Exemplos de Tweets Relacionados aos Possíveis Impactos das Aglomerações no Evento



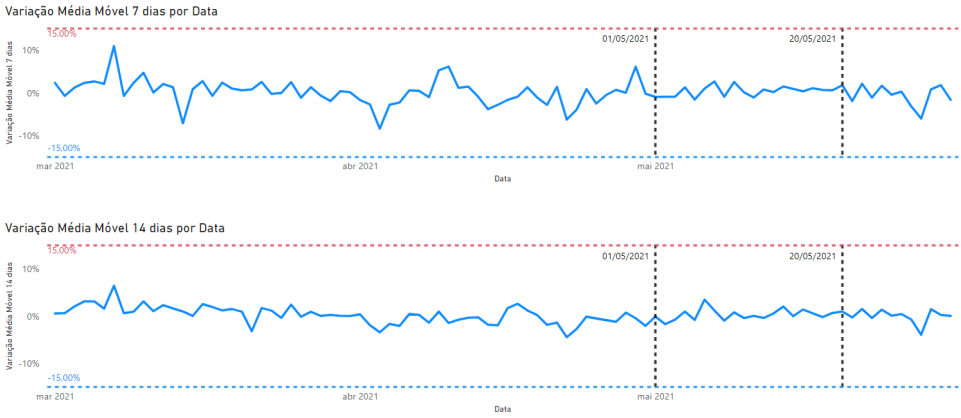
Fonte: Elaborado pelo Autor

Para poder verificar expandir esta análise, foi aplicada a combinação desses resultados com os dados quantitativos obtidos. Inicialmente, é necessário é importante conhecer o cenário geral para então poder analisar os dias próximos ao evento. Por isso, foi desenvolvido um dashboard que traz os dados totais dos casos, óbitos e taxa de letalidade (óbitos/casos), análises ao longo do tempo dos casos e óbitos ilustrados no Apêndice K e da acumulação destes ilustrados no Apêndice L. A partir da análise do cenário geral, a informação a ser analisada foi o número de casos, pois a ocorrência do óbito não apresenta regularidade quanto ao número de dias para este ocorrer. Além disso, foi escolhido a média móvel de 7 dias e a média móvel de 14 dias como

critério de avaliação. Utilizando estas médias, variações positivas ou negativas de até 15% são considerados estáveis, ou seja, para aferir um possível impacto das manifestações foram buscadas variações acimas desse limiar. (SANTOS et al., 2021).

Com todas medidas criadas, ilustradas no Apêndice M, criou-se a análise dos casos frente as médias móveis de 7 e 14 dias, para uma primeira visão e está ilustrada no Apêndice N. Como essa ilustração dificulta a análise de variações, foi desenvolvido um dashboard que utiliza diretamente a variação das médias móveis, assim, pode-se encontrar com maior facilidade como se comportaram os casos no período de análise. Este dashboard está ilustrado na Figura 19. Ao se analisar o mesmo, destaca-se que embora hajam dias que possuem algumas variações maiores perante os demais, todas variações ficaram em até 15%, ou seja, apresentam estabilidade. (SANTOS et al., 2021).

Figura 19 – Dashboard da Variação das Médias Móveis no Período

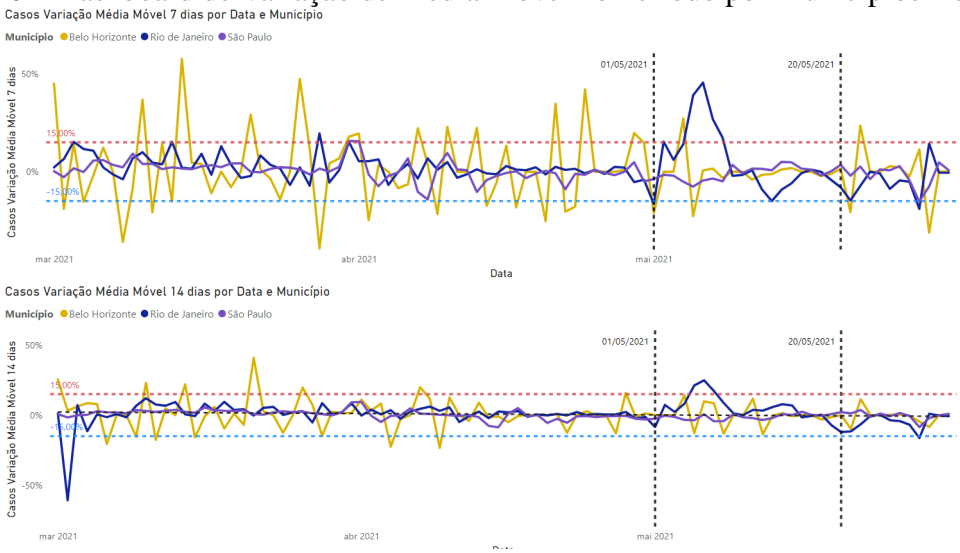


Fonte: Elaborado pelo Autor

Contudo, como as manifestações ocorreram com maior intensidade em alguns municípios, foram escolhidos os três municípios com maiores aglomerações noticiadas para se analisar. (G1, 2021). Os municípios utilizados foram: Belo Horizonte, Rio de Janeiro e São Paulo. A Figura 20 ilustra a análise da variação das médias móveis nestes municípios. Com esta análise mais granular, é possível identificar na variação da média móvel de 7 dias uma instabilidade em Belo Horizonte no dia 04/05/2021 com uma variação de 27,17% e quatro dias de grande alteração no Rio de Janeiro, chegando a variações de 45,46%. Ainda analisando a variação da média móvel de 7 dias, percebe-se uma queda em Belo Horizonte no dia 05/05/2021, indicada provavelmente pela alta do dia anterior, ou seja, a situação deve ser analisada com mais detalhes, mas pode ser apenas uma variação atípica na contagem de casos. Na análise da variação da média móvel de 14 dias encontra-se maior estabilidade, contudo o Rio de Janeiro mantém 3 dias com variações acima de 15%. Isso gera evidências que a cidade do Rio de Janeiro pode ter tido aumentos devidos as manifestações, portanto, cabe uma apuração mais detalhada destes dados. São Paulo mantém estabilidade durante o período analisado.

A utilização da análise qualitativa combinada a análise quantitativa demonstrou os principais impactos da manifestação. Os dados qualitativos conseguiram realizar uma exploração dos da-

Figura 20 – Dashboard de Variação de Média Móvel no Período por Municípios Destacados



Fonte: Elaborado pelo Autor

dos de forma muito positiva, indicando os assuntos mais comentados durante o evento. Um dos assuntos, referente ao possível aumento do número de casos, foi analisado quantitativamente e visto que embora não acha confirmação, existem indícios que as aglomerações ocorridas no evento podem ter causado o aumento do número de contaminado principalmente ao analisar os dados individuais de cada município.

6 CONCLUSÕES E TRABALHOS FUTUROS

O presente artigo analisou o impacto da utilização do thick data combinado ao big data. Isso ocorreu através de um revisão sistemática da literatura, somada a uma bibliometria, etapas desenvolvidas como conscientização do problema durante a aplicação do método Design Science Research. O DSR possibilitou o entendimento das lacunas e possibilidades de melhorias da literatura, a geração dos requisitos que foram derivados em design principles afim de compor o modelo conceitual proposto neste trabalho. Este modelo visa contribuir para o entendimento da necessidade da utilização de dados qualitativos junto aos quantitativos. Através dessa estrutura, será possível melhorar a utilização do big data, utilizando este juntamente ao thick data. Além disso, a aplicação de técnicas qualitativas tendem a realizar análises mais completas e, portanto, extrair mais informações de estruturas de big data.

O trabalho apresentou uma resposta favorável ao uso do thick data para o desenvolvimento de análises mais assertivas junto a uma estrutura de big data. Isso foi visto tanto na análise de adoção do modelo frente aos cases observados na literatura que demonstraram diversas aplicações do thick data a inúmeras áreas, quanto no desenvolvimento do case prático. O case prático conseguiu, além de compreender os impactos sociais do evento analisado através da análise qualitativa, corroborar com a expectativas dos autores dos tweets com a posterior análise quan-

titativa aplicada.

Com os novos resultados, há possibilidade dos números referentes a eficácia e rentabilidade de projetos de big data melhorarem, e, principalmente, dos resultados destes projetos beneficiarem diversas áreas de aplicação, tornando-se assim uma contribuição do presente trabalho. Além disso, também foi reforçado o impacto do uso do thick data, reforçando a importância da combinação de análises qualitativas e quantitativas.

Dentre as limitações do trabalho, destaca-se a possibilidade de uma pesquisa mais extensa como critério de validação do modelo. Desta forma, novos trabalhos podem evoluir e testar o modelo proposto. Complementar a isto, o modelo buscou trazer a importância do thick data dentro de uma estrutura de dados de big data, contudo ele ainda pode ser corretamente incluído dentro do fluxo dos dados e compreendido no contexto geral da análise de dados.

Como um possível trabalho futuro, pode-se gerar uma estrutura que contemple tanto o modelo de dados quanto o fluxo de trabalho. Assim como, o desenvolvimento de um case prático que utilize outras técnicas etnográficas, como a análise de grupos de usuários que pode ser feita através de banco de dados orientados a grafos, e também a utilização de um case onde haja a necessidade de lidar com dados em tempo real.

Referências

ALAOUI, O. Y. et al. CREATING STRATEGIC BUSINESS VALUE from BIG DATA ANALYSIS - APPLICATION TELECOM NETWORK DATA and PLANNING DOCUMENTS. In: **International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives**. [S.l.]: International Society for Photogrammetry and Remote Sensing, 2019. v. 42, n. 4/W16, p. 691–695. ISSN 16821750.

ASAY, M. **8 Reasons Big Data Projects Fail**. 2014. Disponível em: <https://www.informationweek.com/big-data/big-data-analytics/8-reasons-big-data-projects-fail/a/d-id/1297842>.

BARDIN, L.; RETO, L. A. **Análise de conteúdo**. [S.l.]: Editora São Paulo, 2016. 279 p. ISBN 9788562938047.

BI, P. **Power BI**. 2021. Disponível em: <https://powerbi.microsoft.com/pt-br>.

BOELLSTORFF, T. Making big data, in theory. **First Monday**, v. 18, n. 10, 2013.

BORNAKKE, T.; DUE, B. L. Big–Thick Blending: A method for mixing analytical insights from big and thick data sources. **Big Data and Society**, v. 5, n. 1, p. 1–16, 2018. ISSN 20539517.

BRASILEIRO, G. F. Coronavírus Brasil. 2021. Disponível em: <https://covid.saude.gov.br>.

CENTER, A. A. **Big data architectures**. 2018. Disponível em: <https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data>.

CHEN, M.; MAO, S.; LIU, Y. Big data: A survey. **Mobile Networks and Applications**, v. 19, n. 2, p. 171–209, 2014. ISSN 1383469X.

DEMCHENKO, Y. et al. Defining architecture components of the Big Data Ecosystem. **International Conference on IEE**, p. 104:112, 2014. Disponível em: <http://www.uazone.org/demch/papers/bddac2014-bd-ecosystem-archi-v05.pdf>.

FIAIDHI, J. Envisioning Insight-Driven Learning Based on Thick Data Analytics with Focus on Healthcare. **IEEE Access**, Institute of Electrical and Electronics Engineers Inc., v. 8, p. 114998–115004, 2020. ISSN 21693536.

G1. **Dia do Trabalho: 1º de maio tem atos contra e a favor do governo**. 2021.

HENDERSON, M.; SHURVILLE, S.; FERNSTROM, K. The quantitative crunch: The impact of bibliometric research quality assessment exercises on academic development at small conferences. **Campus-Wide Information Systems**, v. 26, n. 3, p. 149–167, 6 2009. ISSN 10650741.

HEVNER, A. R. et al. Design science in information systems research. **MIS Quarterly: Management Information Systems**, Management Information Systems Research Center, v. 28, n. 1, p. 75–105, 2004. ISSN 02767783.

KHALID, B. Big Data in Economic Analysis : Advantages and Challenges. **Ijsser.Org**, n. August, p. 5196–5204, 2019. ISSN 2455-8834. Disponível em: <www.ijsser.org>.

KIBBLE, R. **Introduction to natural language processing**. [S.l.], 2013. Disponível em: <www.londoninternational.ac.uk>.

KUECHLER, B.; VAISHNAVI, V. **On Theory development in design science research: Anatomy of a research project**. 2008. 1–15 p.

MARWICK, A. E. Ethnographic and qualitative research on Twitter. **Twitter and Society**, p. 109–122, 2014.

METH, H.; MUELLER, B.; MAEDCHE, A. Journal of the Association for Information Designing a Requirement Mining System. **Journal of the Association for Information Systems**, v. 16, n. 9, p. 799–837, 2015.

MEULEN, R. v. d. **Gartner Survey Reveals Investment in Big Data Is Up but Fewer Organizations Plan to Invest**. 2016. Disponível em: <https://www.gartner.com/en/newsroom/press-releases/2016-10-04-gartner-survey-reveals-investment-in-big-data-is-up-but-fewer-organizations-plan-to-invest>

MOHAMMED, S.; FIAIDHI, J. **Thick Data: A New Qualitative Analytics for Identifying Customer Insights**. [S.l.]: IEEE Computer Society, 2019. 4–13 p.

MOHER, D. et al. **Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement**. 2009.

MongoDB. **MongoDB**. 2021. Disponível em: <https://www.mongodb.com>.

NLTK. **NTLK Documentation**. 2021. Disponível em: <http://www.nltk.org/howto/portuguese_en.html>.

PINK, S. et al. **Digital Ethnography: Principles and Practice**. [S.l.]: SAGE Publications Ltd, 2016. 202 p. ISBN 978-1-4739-0237-4.

REP, B. **Competitive edge that thick data can bring to your business**. 2017. Disponível em: <https://busitelce.com/competitive-edge-that-thick-data-can-bring-to-your-business>.

RIETSCHE, R. et al. Design and evaluation of an It-based formative feedback tool to foster student performance. **International Conference on Information Systems 2018, ICIS 2018**, n. Riley 2004, p. 1–17, 2018.

SANTOS, P. O. O. et al. **AFINAL, PARA QUE SERVE A MÉDIA MÓVEL?** 2021. Disponível em: <https://coronavirus.saude.mg.gov.br/blog/138-media-movel>.

SINGH, A.; AHMAD, S. Architecture of Data Lake. **International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2019 IJSRCSEIT** |, v. 5, n. 2, p. 2456–3307, 2019. Disponível em: <https://doi.org/10.32628/CSEIT1952121>.

SMETS, A.; LIEVENS, B. Human Sensemaking in the Smart City: A Research Approach Merging Big and Thick Data. **Ethnographic Praxis in Industry Conference Proceedings**, v. 2018, n. 1, p. 179–194, 2018.

Twitter API. Twitter API. 2021. Disponível em: <https://developer.twitter.com/en/docs/twitter-api>.

WANG, T. **Why Big Data Needs Thick Data**. 2016. Disponível em: <https://medium.com/ethnography-matters/why-big-data-needs-thick-data-b4b3e75e3d7>.

WANG, T. **TED Talk The human insights missing from big data**. 2017. Disponível em: <https://www.ted.com/talks/tricia_wang_the_human_insights_missing_from_big_data>.

APÊNDICE A – LIMITAÇÃO 7 DIAS DE BUSCA NA API DO TWITTER

Twitter data is used by developers, students, and researchers to study various topics. You may want to study the conversation around a topic using hashtags, set of keywords, etc. from the last few days. For example, maybe you want to study the sentiment of Tweets about a football game over the weekend. You can get data for such use cases using Twitter's recent search endpoint. This endpoint gives you conversations from Twitter for the last 7 days.

Fonte: Disponível na documentação oficial da API do Twitter (Twitter API, 2021)

APÊNDICE B – LIMITAÇÃO DE MÁXIMO DE 100 TWEETS

`max_results` OPTIONAL	integer	The maximum number of search results to be returned by a request. A number between 10 and the system limit (currently 100). By default, a request response will return 10 results.

Fonte: Disponível na documentação oficial da API do Twitter (Twitter API, 2021)

APÊNDICE C – ARRAY DE HORÁRIOS PARA OBTENÇÃO DOS DADOS

```
times_to_search = [
    "2021-04-29T00:00:00Z", "2021-04-29T01:00:00Z", "2021-04-29T02:00:00Z", "2021-04-29T03:00:00Z", "2021-04-29T04:00:00Z", "2021-04-29T05:00:00Z",  "2021-04-29T06:00:00Z", "2021-04-29T07:00:00Z",
    "2021-04-29T08:00:00Z", "2021-04-29T09:00:00Z", "2021-04-29T10:00:00Z", "2021-04-29T11:00:00Z", "2021-04-29T12:00:00Z", "2021-04-29T13:00:00Z", "2021-04-29T14:00:00Z", "2021-04-29T15:00:00Z",
    "2021-04-29T16:00:00Z", "2021-04-29T17:00:00Z", "2021-04-29T18:00:00Z", "2021-04-29T19:00:00Z", "2021-04-29T20:00:00Z", "2021-04-29T21:00:00Z", "2021-04-29T22:00:00Z", "2021-04-29T23:00:00Z",

    "2021-04-30T00:00:00Z", "2021-04-30T01:00:00Z", "2021-04-30T02:00:00Z", "2021-04-30T03:00:00Z", "2021-04-30T04:00:00Z", "2021-04-30T05:00:00Z",  "2021-04-30T06:00:00Z", "2021-04-30T07:00:00Z",
    "2021-04-30T08:00:00Z", "2021-04-30T09:00:00Z", "2021-04-30T10:00:00Z", "2021-04-30T11:00:00Z", "2021-04-30T12:00:00Z", "2021-04-30T13:00:00Z", "2021-04-30T14:00:00Z", "2021-04-30T15:00:00Z",
    "2021-04-30T16:00:00Z", "2021-04-30T17:00:00Z", "2021-04-30T18:00:00Z", "2021-04-30T19:00:00Z", "2021-04-30T20:00:00Z", "2021-04-30T21:00:00Z", "2021-04-30T22:00:00Z", "2021-04-30T23:00:00Z",

    "2021-05-01T00:00:00Z", "2021-05-01T01:00:00Z", "2021-05-01T02:00:00Z", "2021-05-01T03:00:00Z", "2021-05-01T04:00:00Z", "2021-05-01T05:00:00Z",  "2021-05-01T06:00:00Z", "2021-05-01T07:00:00Z",
    "2021-05-01T08:00:00Z", "2021-05-01T09:00:00Z", "2021-05-01T10:00:00Z", "2021-05-01T11:00:00Z", "2021-05-01T12:00:00Z", "2021-05-01T13:00:00Z", "2021-05-01T14:00:00Z", "2021-05-01T15:00:00Z",
    "2021-05-01T16:00:00Z", "2021-05-01T17:00:00Z", "2021-05-01T18:00:00Z", "2021-05-01T19:00:00Z", "2021-05-01T20:00:00Z", "2021-05-01T21:00:00Z", "2021-05-01T22:00:00Z", "2021-05-01T23:00:00Z",

    "2021-05-02T00:00:00Z", "2021-05-02T01:00:00Z", "2021-05-02T02:00:00Z", "2021-05-02T03:00:00Z", "2021-05-02T04:00:00Z", "2021-05-02T05:00:00Z",  "2021-05-02T06:00:00Z", "2021-05-02T07:00:00Z",
    "2021-05-02T08:00:00Z", "2021-05-02T09:00:00Z", "2021-05-02T10:00:00Z", "2021-05-02T11:00:00Z", "2021-05-02T12:00:00Z", "2021-05-02T13:00:00Z", "2021-05-02T14:00:00Z", "2021-05-02T15:00:00Z",
    "2021-05-02T16:00:00Z", "2021-05-02T17:00:00Z", "2021-05-02T18:00:00Z", "2021-05-02T19:00:00Z", "2021-05-02T20:00:00Z", "2021-05-02T21:00:00Z", "2021-05-02T22:00:00Z", "2021-05-02T23:00:00Z",

    "2021-05-03T00:00:00Z", "2021-05-03T01:00:00Z", "2021-05-03T02:00:00Z", "2021-05-03T03:00:00Z", "2021-05-03T04:00:00Z", "2021-05-03T05:00:00Z",  "2021-05-03T06:00:00Z", "2021-05-03T07:00:00Z",
    "2021-05-03T08:00:00Z", "2021-05-03T09:00:00Z", "2021-05-03T10:00:00Z", "2021-05-03T11:00:00Z", "2021-05-03T12:00:00Z", "2021-05-03T13:00:00Z", "2021-05-03T14:00:00Z", "2021-05-03T15:00:00Z",
    "2021-05-03T16:00:00Z", "2021-05-03T17:00:00Z", "2021-05-03T18:00:00Z", "2021-05-03T19:00:00Z", "2021-05-03T20:00:00Z", "2021-05-03T21:00:00Z", "2021-05-03T22:00:00Z", "2021-05-03T23:00:00Z"
    ]
```

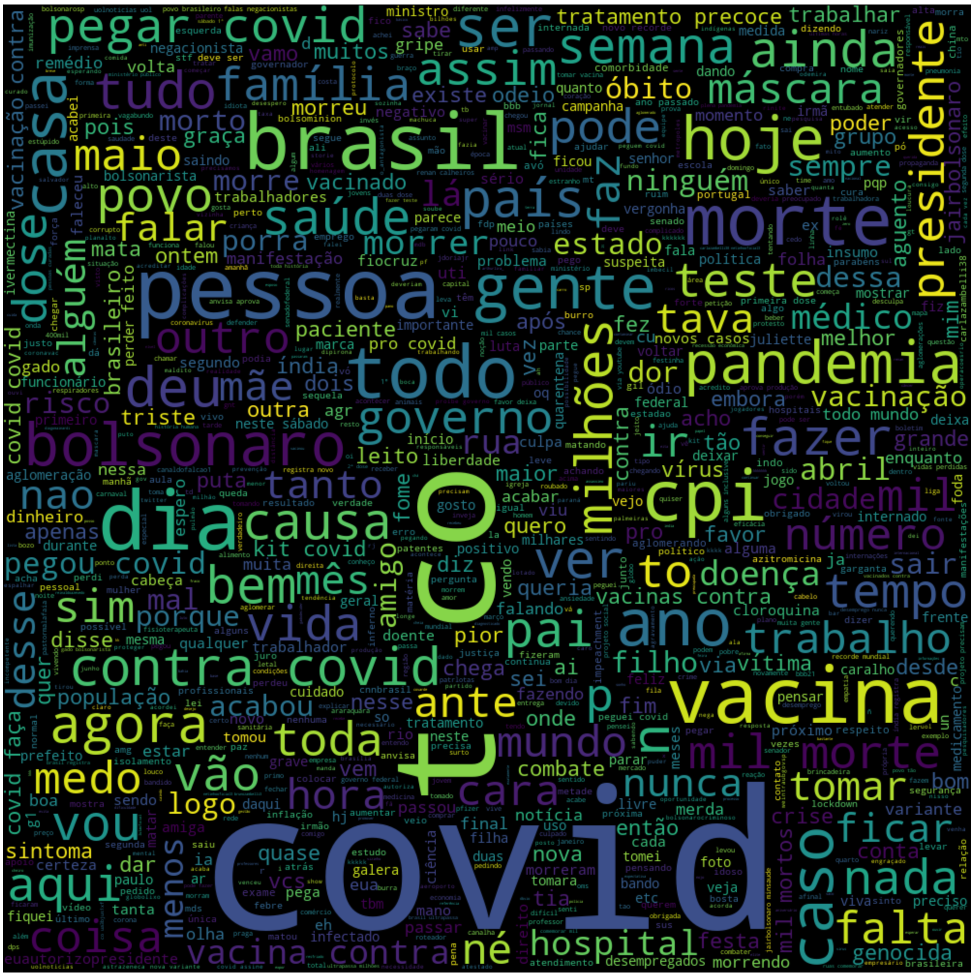
Fonte: Elaborado pelo Autor

APÊNDICE D – LISTA DE PALAVRAS REMOVIDAS DA ANÁLISE

```
from string import punctuation
import nltk
stopwords = set(nltk.corpus.stopwords.words('portuguese') + list(punctuation) + list(["https", "//t", "pra", "q", "...", "vai", "tá", "ter", "pq", "sobre", "vc", "tô", "aí", "ta", "","“", "”"]))
```

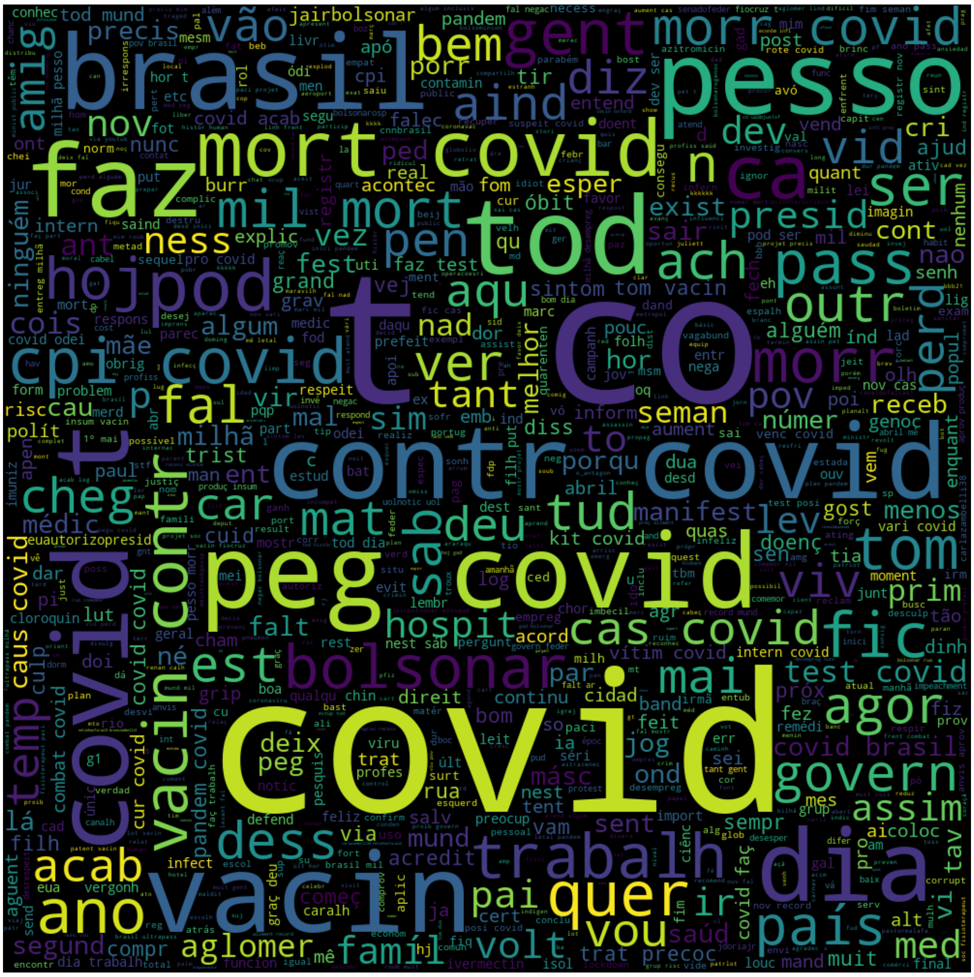
Fonte: Elaborado pelo Autor

APÊNDICE E – WORD CLOUD DAS PALAVRAS ANTES DO EVENTO



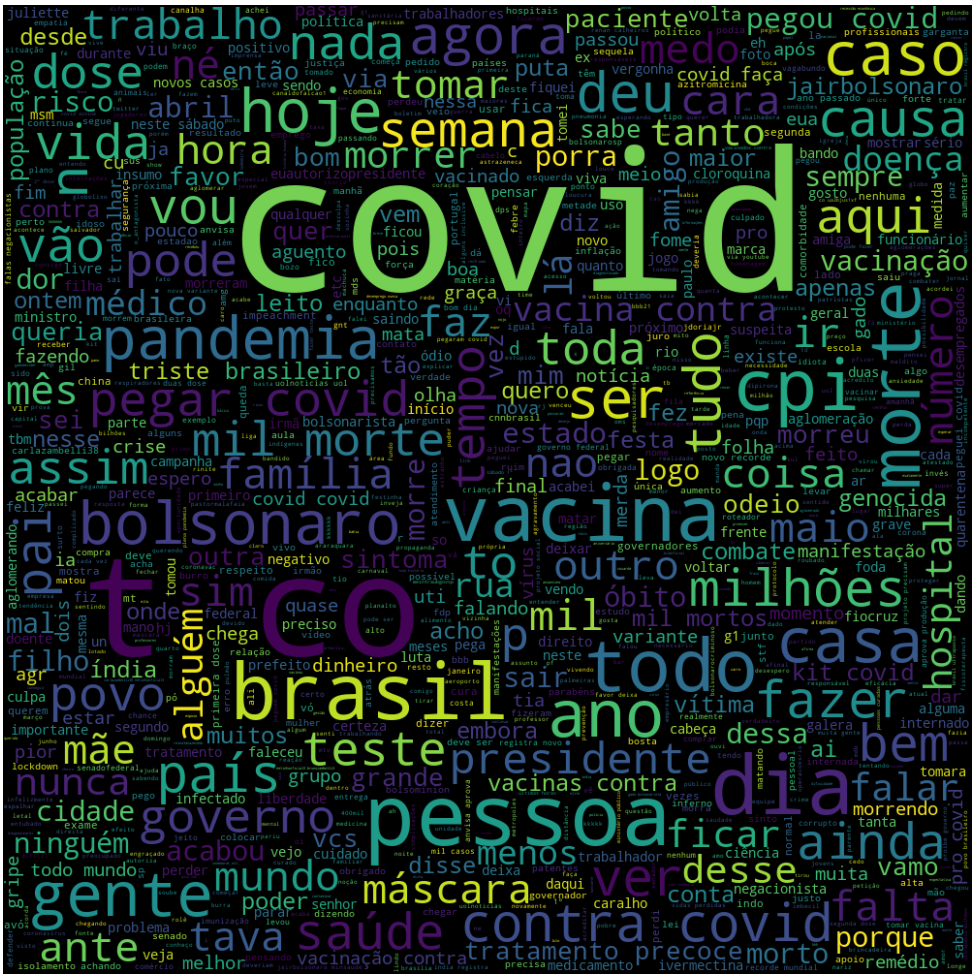
Fonte: Elaborado pelo Autor

APÊNDICE F – WORD CLOUD DOS STEMMINGS ANTES DO EVENTO



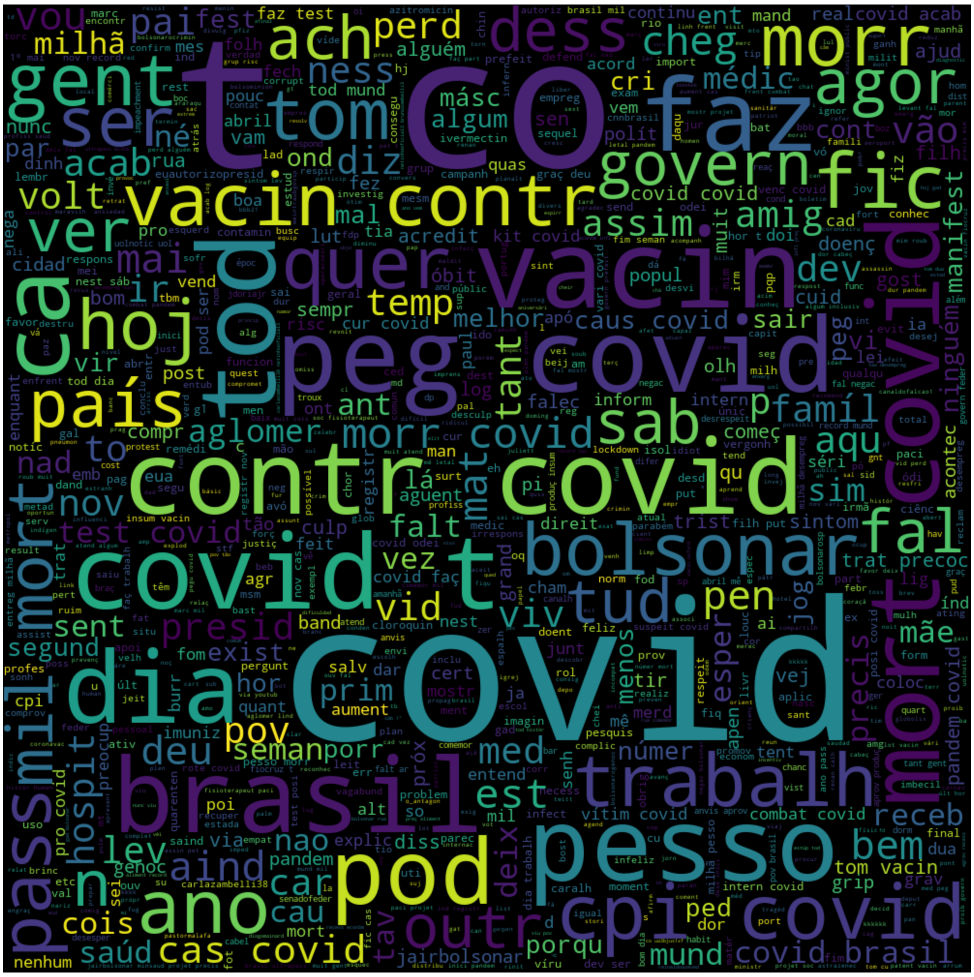
Fonte: Elaborado pelo Autor

APÊNDICE G – WORD CLOUD DAS PALAVRAS DURANTE O EVENTO



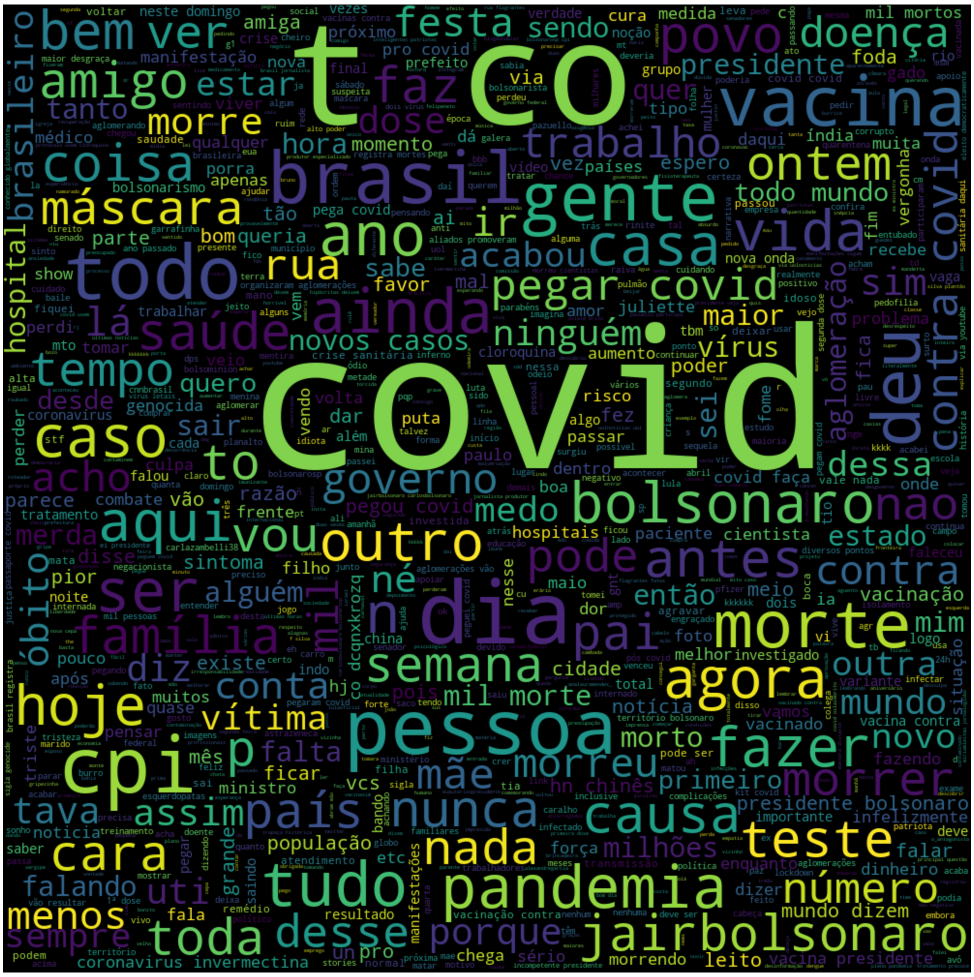
Fonte: Elaborado pelo Autor

APÊNDICE H – WORD CLOUD DOS STEMMINGS DURANTE O EVENTO



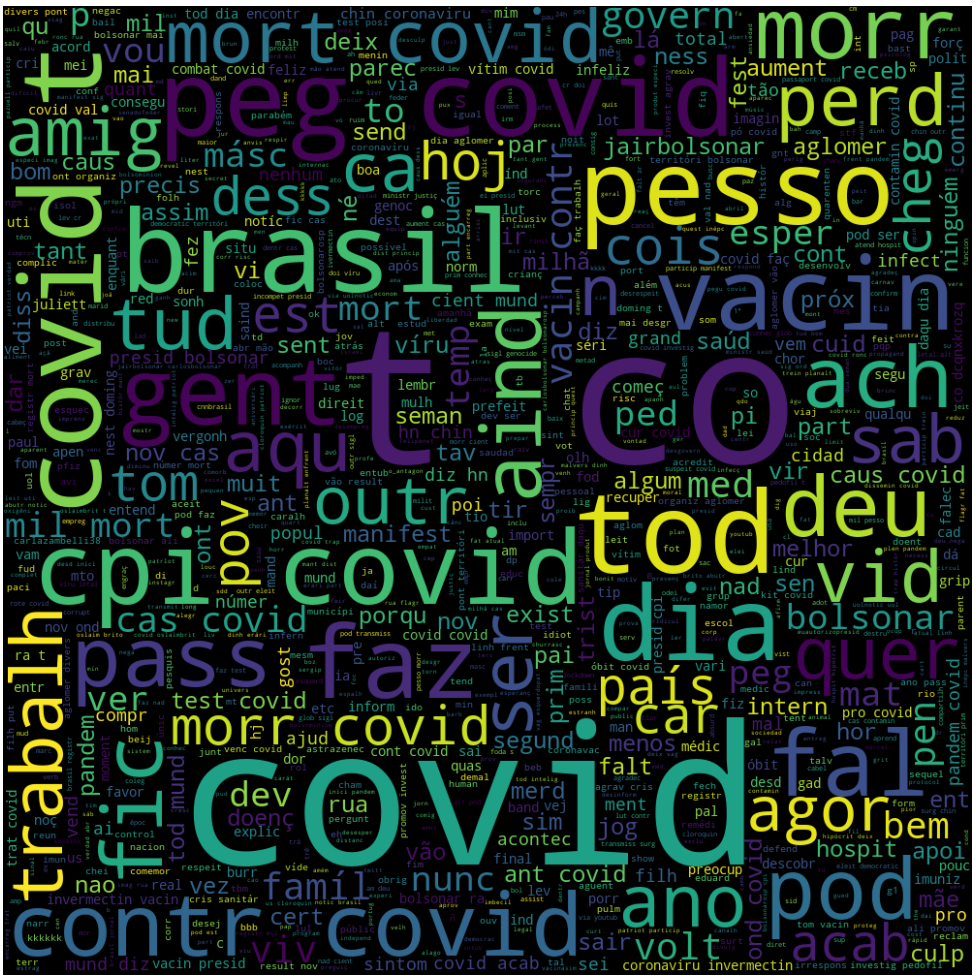
Fonte: Elaborado pelo Autor

APÊNDICE I – WORD CLOUD DAS PALAVRAS DEPOIS DO EVENTO



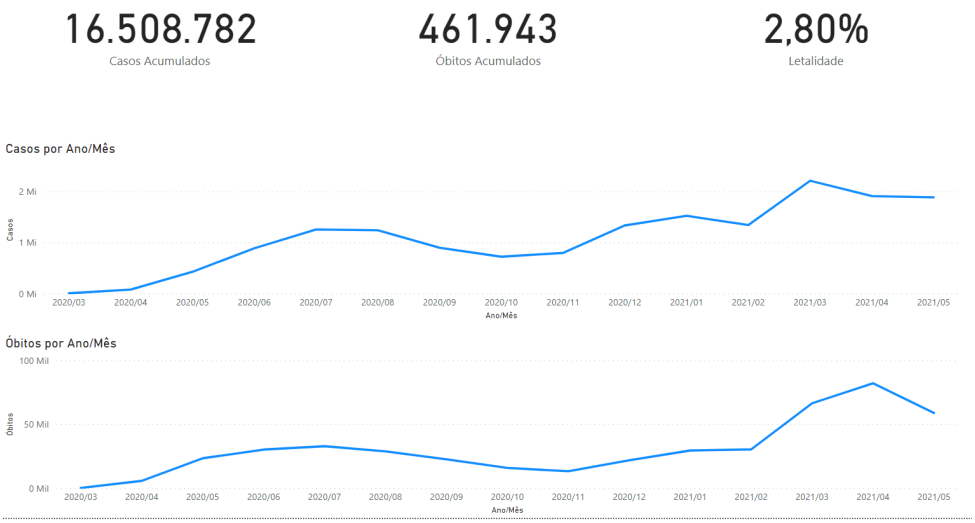
Fonte: Elaborado pelo Autor

APÊNDICE J – WORD CLOUD DOS STEMMINGS DEPOIS DO EVENTO



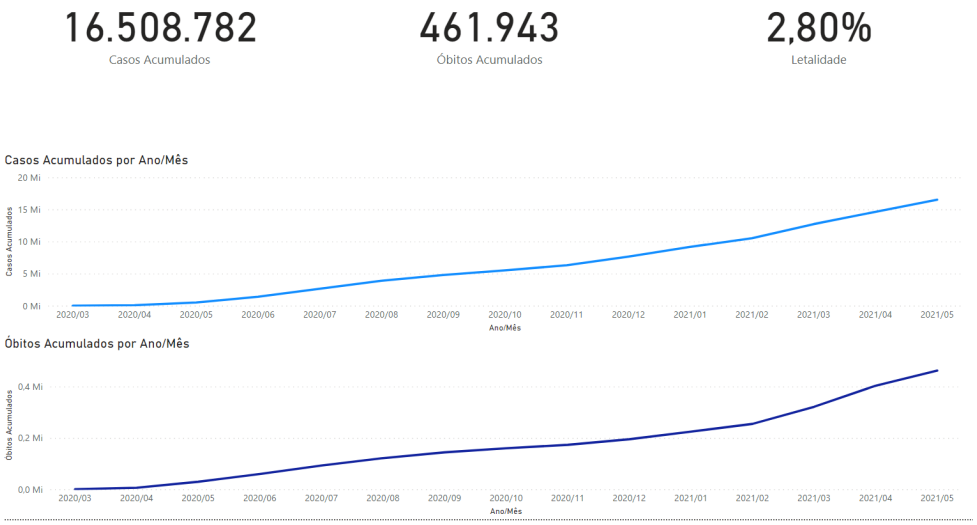
Fonte: Elaborado pelo Autor

APÊNDICE K – DASHBOARD DADOS GERAIS



Fonte: Elaborado pelo Autor

APÊNDICE L – DASHBOARD DADOS GERAIS ACUMULADOS



Fonte: Elaborado pelo Autor

APÊNDICE M – MEDIDAS DESENVOLVIDAS NO POWER BI

```
Casos = SUM(Dados[Casos])

Casos Acumulados =
CALCULATE(
    SUM( Dados[Casos]),
    FILTER(ALL('Calendário'),  'Calendário'[Data] <= MAX('Calendário'[Data])
))


Óbitos = SUM(Dados[Óbitos])

Óbitos Acumulados =
CALCULATE(
    SUM( Dados[Óbitos]),
    FILTER(ALL('Calendário'),  'Calendário'[Data] <= MAX('Calendário'[Data])
))


Letalidade = [Óbitos Acumulados] / [Casos Acumulados]

Casos Variação Média Móvel 7 dias =
CALCULATE(
    AVERAGE('Dados'[Casos]),
    FILTER(ALL('Calendário'), 'Calendário'[Data] > MAX('Calendário'[Data]) - 7 && 'Calendário'[Data] <= MAX
('Calendário'[Data]))
)
/
CALCULATE(
    AVERAGE('Dados'[Casos]),
    FILTER(ALL('Calendário'), 'Calendário'[Data] > MAX('Calendário'[Data]) - 8 && 'Calendário'[Data] <= MAX
('Calendário'[Data])-1)
) - 1


Casos por Data = SUMMARIZE(Dados, 'Dados'[Data], "Casos por Data", SUM(Dados[Casos]))

Casos Média Móvel 7 dias Data =
CALCULATE(
    AVERAGE('Casos por Data'[Casos por Data]),
    FILTER(ALL('Calendário'), 'Calendário'[Data] > MAX('Calendário'[Data]) - 7 && 'Calendário'[Data] <= MAX
('Calendário'[Data]))
)

Casos Variação Média Móvel 7 dias por Data =
CALCULATE(
    AVERAGE('Casos por Data'[Casos por Data]),
    FILTER(ALL('Calendário'), 'Calendário'[Data] > MAX('Calendário'[Data]) - 7 && 'Calendário'[Data] <= MAX
('Calendário'[Data]))
)
/
CALCULATE(
    AVERAGE('Casos por Data'[Casos por Data]),
    FILTER(ALL('Calendário'), 'Calendário'[Data] > MAX('Calendário'[Data]) - 8 && 'Calendário'[Data] <= MAX
('Calendário'[Data])-1)
) - 1
```

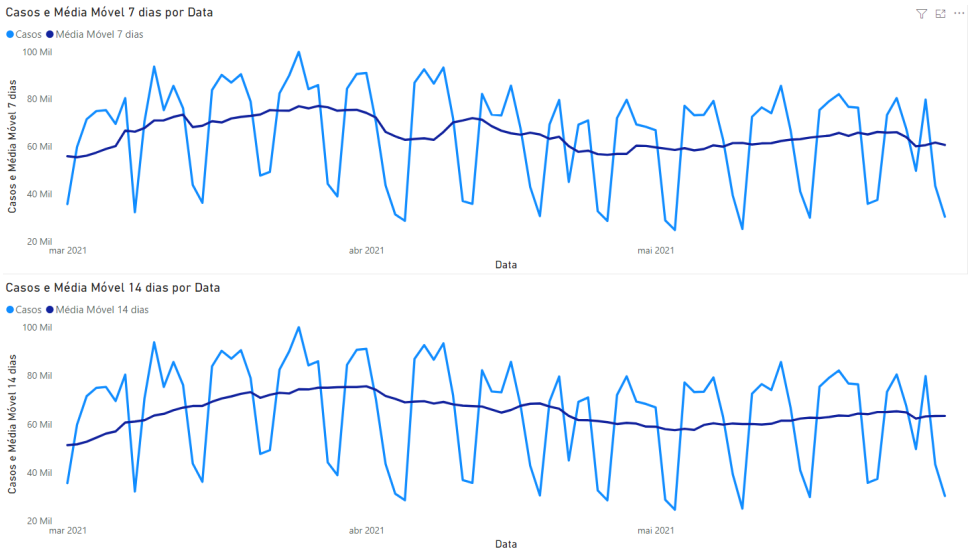
```
Casos Variação Média Móvel 14 dias =
CALCULATE(
    AVERAGE('Dados'[Casos]),
    FILTER(ALL('Calendário'), 'Calendário'[Data] > MAX('Calendário'[Data]) - 14 && 'Calendário'[Data] <= MAX
('Calendário'[Data]))
)
/
CALCULATE(
    AVERAGE('Dados'[Casos]),
    FILTER(ALL('Calendário'), 'Calendário'[Data] > MAX('Calendário'[Data]) - 15 && 'Calendário'[Data] <= MAX
('Calendário'[Data])-1)
) - 1


Casos Média Móvel 14 dias Data =
CALCULATE(
    AVERAGE('Casos por Data'[Casos por Data]),
    FILTER(ALL('Calendário'), 'Calendário'[Data] > MAX('Calendário'[Data]) - 14 && 'Calendário'[Data] <= MAX
('Calendário'[Data]))
)

Casos Variação Média Móvel 14 dias por Data =
CALCULATE(
    AVERAGE('Casos por Data'[Casos por Data]),
    FILTER(ALL('Calendário'), 'Calendário'[Data] > MAX('Calendário'[Data]) - 14 && 'Calendário'[Data] <= MAX
('Calendário'[Data]))
)
/
CALCULATE(
    AVERAGE('Casos por Data'[Casos por Data]),
    FILTER(ALL('Calendário'), 'Calendário'[Data] > MAX('Calendário'[Data]) - 15 && 'Calendário'[Data] <= MAX
('Calendário'[Data])-1)
) - 1
```

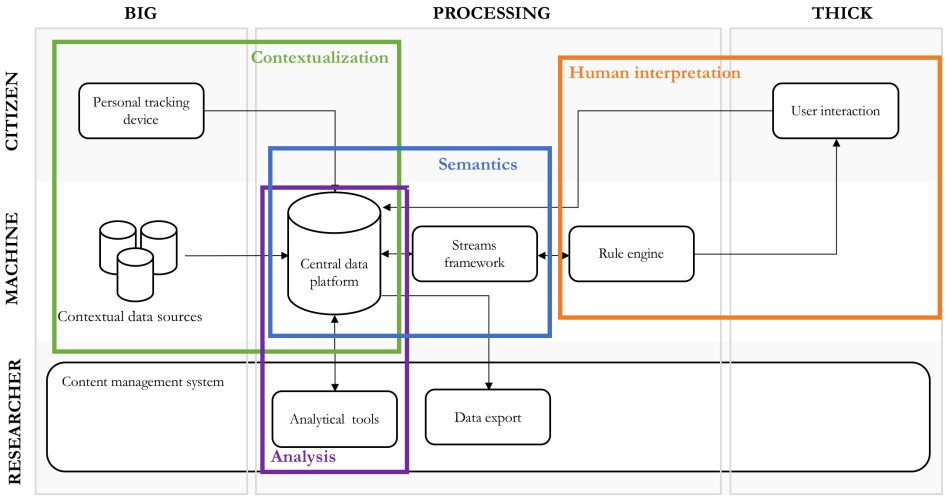
Fonte: Elaborado pelo Autor

APÊNDICE N – DASHBOARD COM CASOS E MÉDIAS MÓVEIS



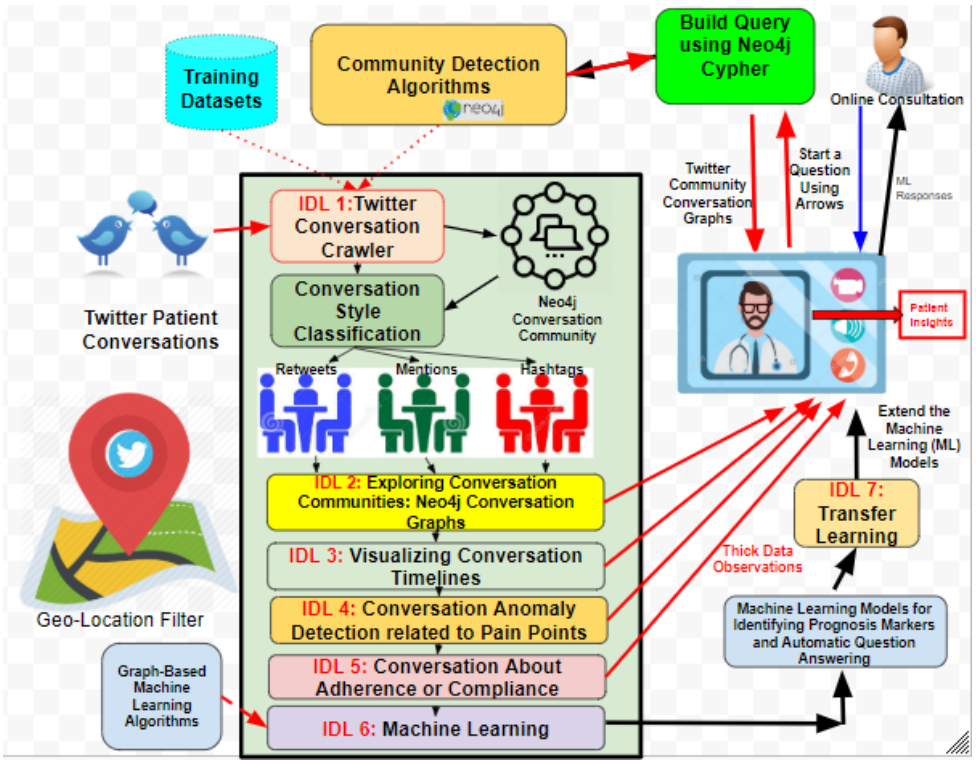
Fonte: Elaborado pelo Autor

ANEXO A – APRESENTAÇÃO ESQUEMÁTICA DO CITIZEN TOOLBOX



Fonte: (SMETS; LIEVENS, 2018)

ANEXO B – MÉTODOS GERAIS DE APRENDIZAGEM ORIENTADOS PELO THICK DATA INSIGHT



Fonte: (FIAIDHI, 2020)