UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE BIBLIOTECONOMIA E COMUNICAÇÃO
DEPARTAMENTO DE COMUNICAÇÃO
CURSO DE JORNALISMO

CAMILA FONTES PESSÔA

TRANSPARÊNCIA E JORNALISMO DE DADOS:

uma análise de reportagens e repositórios no GitHub mantidos por redações brasileiras

Porto Alegre
2022

CAMILA FONTES PESSÔA

TRANSPARÊNCIA E JORNALISMO DE DADOS:

uma análise de reportagens e repositórios no GitHub mantidos por redações brasileiras

Trabalho de Conclusão de Curso apresentado à Faculdade de Biblioteconomia e Comunicação da Universidade Federal do Rio Grande do Sul como requisito parcial à obtenção do grau de Bacharel em Jornalismo.

Orientador: Prof. Dr. Marcelo Ruschel Träsel

Porto Alegre

2022

AGRADECIMENTOS

Não é possível chegar a lugar nenhum sozinho. Ao longo do caminho, felizmente tive inúmeras pessoas que me apoiaram e de alguma forma me ajudaram a chegar até aqui e a produzir este trabalho.

Por isso, agradeço ao meu pai, Sávio, que desde sempre foi também um professor para mim, inclusive nesta monografia, ajudando-me a ter ideias, construir e avaliar minhas análises. À minha mãe, Celda, pelo carinho, atenção e ensinamentos. Também aos meus irmãos, Beatriz e Gabriel, por me fazerem rir em momentos de tensão e me compreenderem de uma forma tão profunda, e a Rosa, que acompanhou minha vida desde o início.

Ao meu orientador, Marcelo Träsel, por seu apoio fundamental e influência na escolha do tema para esta pesquisa, ao meu examinador, Felipe, que esteve me acompanhando e me ensinando muito nesses últimos semestres, e à minha examinadora, Marília, por ter aceitado participar da banca mesmo estando em outro fuso horário, no extremo hemisfério norte do mundo. Também agradeço à UFRGS, à Nossa Escola, e a todos os professores que fizeram parte da minha caminhada.

Aos veículos e jornalistas que disponibilizaram seus trabalhos no GitHub, sem os quais esta pesquisa não existiria, e ao jornalista de GZH Marcel Hartmann, por ter me ajudado com informações para este trabalho.

À minha família em Porto Alegre: meus tios, Roberto, Telma, Pedro e Zinha, também ao Jefferson, à Patrícia e a todos os outros familiares que me acolheram durante a graduação. Sem falar dos avós, tios e primos que me acompanharam de Aracaju e de outras partes do Brasil e do mundo.

À Irene, Dariva e Tiago, amigos de longa data que me ofereceram um suporte emocional enorme durante todos esses anos, à Isadora e à Nicolle, amigas e parceiras em todas as atividades da graduação, e a todos os amigos, colegas de faculdade e de trabalho e outras pessoas que passaram por esse caminho.

Por fim, um agradecimento especial à Mary, à Carol e ao Victor, que acompanharam o meu dia a dia durante toda a produção do TCC e ouviram minhas angústias.

RESUMO

O objetivo desta monografia foi entender o uso do GitHub como ferramenta para transparência no jornalismo e os métodos utilizados por jornalistas-programadores brasileiros para obter e estruturar informações. A metodologia aplicada foi o modelo de estudo de caso por ilustração desenvolvido pelo Grupo de Pesquisa em Jornalismo On-line (GJOL), voltada para o jornalismo digital. O *corpus* é composto por dez peças jornalísticas e pelos repositórios no GitHub que explicam os métodos utilizados na produção de cada uma delas. Ao fim da análise conclui-se que há uma evidente preocupação com o entendimento dos métodos pelo público na maioria das peças, que existe uma variedade de abordagens aplicadas, a mais comum sendo a limpeza de dados, e que o GitHub é uma ferramenta com potencial para ampliar a transparência no jornalismo.

Palavras-chave: transparência; jornalismo de dados; dados abertos; código aberto; programação.

ABSTRACT

This monograph aimed to understand the use of GitHub as a tool for transparency in journalism and the methods used by programmer-journalists to obtain and structure information. The methodology applied was a model of illustration case study developed by the Research Group on Online Journalism (GJOL) focusing on digital journalism. The corpus is composed of ten journalistic pieces and the GitHub repositories that explain the methods used in the production of each one of them. At the end of this analysis it is concluded that there is an evident concern with the audience's understanding of the methods in most of the pieces, that a variety of approaches are applied, the most common being data cleaning, and that GitHub has a potential to increase transparency in journalism.

Keywords: transparency; programmer-journalist; open data; open source; programming.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

Em um mundo capitalista onde o lucro e a propriedade parecem ser sempre as prioridades, pessoas e organizações que compartilham com todos, de forma gratuita e aberta, o fruto de seu árduo trabalho são um respiro. Mesmo que hoje estejam evidentes os males que a internet e outras tecnologias podem trazer, como a propagação de desinformação, ascensão de grupos extremistas e controle de informações pessoais de usuários por empresas, as oportunidades de democratização do conhecimento, trabalho em conjunto e busca de soluções trazidas pela comunicação em rede são inúmeras. Ainda mais no contexto atual, em que há uma quantidade imensa de informações acessíveis, apenas esperando olhares atentos e tecnicamente capacitados para explorá-las.

A vontade de desvendar essas ferramentas que tornam os dados mais acessíveis foi o que motivou a escolha do tema para este trabalho. Entender como se caracteriza uma forma de fazer jornalismo ainda muito nova é um dos objetivos.

Dessa forma, o questionamento principal deste trabalho é: **como os jornalistas-programadores brasileiros exercem a transparência sobre os métodos que utilizam para obter informações a partir de bases de dados?**

Assim, o objetivo geral é descrever, a partir da análise de repositórios de dados e *scripts* vinculados a reportagens, as estratégias de transparência metodológica adotadas por redações brasileiras que publicam reportagens com uso de linguagens de programação. E, considerando essas questões, os objetivos específicos foram a) identificar reportagens com uso de linguagens de programação realizadas por redações brasileiras; b) descrever como os dados e procedimentos metodológicos são divulgados pelas redações na plataforma GitHub; c) identificar os principais tipos de operações computacionais usados em um conjunto de jornais brasileiros.

Para essas análises, foi aplicada a metodologia de estudo de caso por ilustração do Grupo de Pesquisa em Jornalismo On-line (GJOL), de pesquisa quantitativa e qualitativa complementadas por estudo de caso. O objeto empírico são dez peças jornalísticas que utilizam dados processados de alguma forma e tiveram os códigos-fonte dos procedimentos computacionais compartilhados no GitHub.

O GitHub é uma plataforma colaborativa de compartilhamento de código-fonte, construída a partir do sistema Git, concebido para gerenciar diferentes

versões de um documento de software. Na plataforma, o código é armazenado em nuvem e pode ser encontrado por qualquer pessoa. O GitHub também tem características de redes sociais, como a presença de perfis de usuários com descrições pessoais e atividade recente, a opção de seguir esses perfis e a presença de um feed com informações de interesse do usuário (BLINCOE et al., 2015).

Com o intuito de explorar o material publicado por veículos e jornalistas nessa plataforma, a estrutura escolhida para esta monografia consiste em um capítulo de introdução, três capítulos de revisão bibliográfica, um capítulo com detalhamento da metodologia utilizada e análises e um capítulo de considerações finais. Cada elemento da estrutura é descrito a seguir.

O capítulo 2 descreve as novas ferramentas e possibilidades que a internet trouxe para a prática jornalística. São discutidos o contexto atual de grande disponibilidade de informações e os conceitos de dados abertos, Lei de Acesso a Informação, Jornalismo Guiado por Dados (JGD) e visualização de dados.

O capítulo 3 traz a discussão para o JGD em específico, estudando como ele pode contribuir para uma maior transparência e objetividade no jornalismo. A adesão de jornalistas de dados à cultura hacker e de código aberto, que tiveram origem entre programadores, também é discutida.

Enquanto isso, o capítulo 4 fala sobre as principais ferramentas utilizadas para extrair informações úteis em meio a uma imensidão de dados disponíveis na web e sobre o conceito de Big Data.

No capítulo 5 são apresentados o objeto de pesquisa e a metodologia. As peças jornalísticas escolhidas pertencem a dez veículos diferentes e foram selecionadas observando-se a disponibilidade de material no GitHub. São apresentadas descrições dos veículos, das peças e das explicações de metodologia e código-fonte encontradas no GitHub. Por fim, é realizada uma análise com base nas categorias criadas a partir de particularidades do objeto de pesquisa observadas na etapa de análise preliminar deste trabalho: *transparência, profundidade, abertura, relevância* e *interatividade*.

Ao final, no capítulo 6, os objetivos das análises são revisitados, concluindo-se que há uma evidente preocupação com o entendimento dos métodos pelo público na maioria das peças e que a maior parte dos trabalhos utilizam

operações como extrair, limpar e filtrar informações coletadas de bancos de dados públicos.[1]

[1] Vale lembrar que os termos Jornalismo de Dados e Jornalismo Guiado por Dados são utilizados como sinônimos nesta monografia, sendo o segundo termo mais utilizado quando citado por outros autores.

2 O JORNALISMO E A INTERNET

Neste capítulo, discute-se como a internet e as informações às quais ela permite acesso mudaram a configuração do jornalismo e trouxeram a necessidade de novas ferramentas para a investigação de informações presentes numa infinidade de dados que hoje estão disponíveis.

Com as transformações tecnológicas ligadas aos avanços e reestruturação do sistema capitalista desde os anos 1980, que levaram à difusão da internet, o acesso à informação foi cada vez mais facilitado, levando ao surgimento da sociedade da informação. De acordo com Castells (2001), essa sociedade se caracteriza pelo melhor uso possível das tecnologias para lidar com a informação, tornando-a central na atividade humana. De acordo com o autor, as características fundamentais dessa sociedade são a informação como matéria-prima, com o desenvolvimento de tecnologias específicas para lidar com ela, diferentemente de tempos anteriores, em que a informação era utilizada majoritariamente para desenvolver tecnologias; a penetrabilidade das novas tecnologias, que passam a influenciar em todas as esferas da sociedade; uma lógica de redes predominante, aplicável para todos os tipos de relações complexas e processos; a flexibilidade, caracterizada por processos reversíveis, que podem ser modificados, reorganizados e reconfigurados; e o crescimento da convergência de tecnologias, contexto em que os percursos de desenvolvimento tecnológico em diferentes áreas interligam-se, transformando também as categorias de acordo com as quais pensamos processos (CASTELLS, 2000).

Ao se pensar nessas questões, pode-se observar que fatores sociais e características humanas pré-existentes influenciam no desenvolvimento das tecnologias e nas aplicações sociais delas (WERTHEIN, 2000). Assim, o desenvolvimento tecnológico que ocorreu a partir da década de 1970 pode estar ligado à cultura da liberdade, inovação individual e iniciativa empreendedora que vinha se desenvolvendo desde a década de 1960 e que acabou se difundindo pela cultura da nossa sociedade (Castells, 2000).

Considerando esses aspectos, o desenvolvimento desse novo paradigma ocorre em ritmos e atinge níveis diferentes de acordo com a sociedade em que os processos estão ocorrendo, reproduzindo desigualdades de renda e

desenvolvimento industrial e, assim, excluindo determinados grupos (WERTHEIN, 2000).

Pelo lado positivo, o desenvolvimento da sociedade da informação e a penetrabilidade das novas tecnologias que ele acarreta podem dar suporte a iniciativas que têm o objetivo de preparar a sociedade para enfrentar e conhecer as transformações, uma vez que permitem implementar a lógica de redes e modelar resultados da criatividade originada da interação complexa, o que alimenta novos sonhos e expectativas de criação a partir da tecnologia. Além disso, flexibilidade e a capacidade de reconfiguração de sistemas incorporam a ideia de aprendizagem, com maior disponibilidade para executar mudanças, e de adaptação e aperfeiçoamento contínuos, tanto intelectuais quanto técnicos (WERTHEIN, 2000).

Nessa sociedade informatizada, o desenvolvimento tecnológico parece não ter limites e modifica continuamente os processos que afetam a vida individual e coletiva, tendo capacidade de trazer facilidades em relação ao bem-estar, lazer e “acesso rápido, ilimitado e eficiente, ao rico acervo do conhecimento humano” (WERTHEIN, 2000).

Porém, o papel da tecnologia na construção de uma “sociedade do conhecimento” inovadora pode trazer riscos, em especial para os países em desenvolvimento, nos quais a incorporação das novas tecnologias demanda investimentos nas capacidades tecnológicas locais e no desenvolvimento de instituições e os avanços vão depender de como tensões entre culturas e modos de organização social vão ser resolvidas. Nesses lugares, é possível que a introdução das tecnologias dê origem a novas formas de exclusão (MANSELL; WEHN, 1998). Para superar essas forças de exclusão, é necessário tomar medidas para promover o acesso universal à infra-estrutura e aos serviços de informação (WERTHEIN, 2000).

O acesso mais generalizado ao conteúdo e fontes do conhecimento traz desafios. Um deles é aumentar o volume de informações qualificadas e de domínio público disponíveis na internet, nos idiomas de cada população. “Isso envolverá convencer o governo e centros produtores de conhecimento financiados por recursos públicos a tornarem disponíveis ao público as informações produzidas” (WERTHEIN, 2000).

2.1 Dados abertos, LAI e Jornalismo de Dados

Com a percepção desse desafio, foram crescendo ao longo do tempo os movimentos para garantir a disponibilização de dados com a possibilidade de livre utilização, reutilização e redistribuição por todo e qualquer indivíduo, chamados de dados abertos. Esse tipo de informação é caracterizada por ter, no máximo, sua fonte original mencionada e a exigência de que sejam usadas as mesmas licenças no ato de compartilhamento. Assim, a abertura de dados pressupõe o interesse em evitar mecanismos de controle e restrições, para que todos possam ter acesso aos dados (OPEN DEFINITION, 2015).

Um desses movimentos em direção a uma maior quantidade de dados abertos foi a Open Government Initiative, criada a partir de um Memorando sobre Transparência e Governo Aberto assinado pelo ex-presidente dos Estados Unidos Barack Obama no seu primeiro dia de governo, em 2009. Essa política define ações para diminuir a influência de interesses especiais no governo (como os lobistas), tornar as informações sobre os destinos de recursos públicos acessíveis a todos os cidadãos e empoderar o público para influenciar na tomada de decisões. Mais tarde, também foi assinado o Open Government Directive, que exigia dos órgãos governamentais ações imediatas em direção a uma maior transparência, participação e colaboração (THE WHITE HOUSE - PRESIDENT BARACK OBAMA, |201?|).

Observando esses pressupostos, três normas principais estão implícitas ao termo dados abertos. A primeira é a disponibilidade de acesso, que pressupõe que os dados estejam disponíveis por inteiro, de forma conveniente e modificável, por um valor não maior que um custo de reprodução razoável, de preferência que podem ser baixados por meio da internet. A segunda norma é a do reúso e redistribuição, de acordo com a qual a forma como os dados são disponibilizados deve oferecer possibilidade de reutilização, redistribuição e combinação com outras bases de dados. E a terceira é a participação universal, norma que define que todos devem ter capacidade de usar os dados, sem restrições a pessoas, objetivos de uso ou áreas de atuação (OPEN KNOWLEDGE FOUNDATION, 2010).

Ao longo dos anos, novas políticas de abertura de dados foram sendo criadas ao redor do mundo. No Brasil, a constituição de 1988 garante o acesso à informação pública e, a partir dela, legislações que previam a divulgação de dados como

orçamentos, outras informações financeiras e atos administrativos começaram a aparecer. A mais detalhada delas é a Lei de Acesso à Informação (LAI), “que regulamentou em detalhes o acesso como regra e o sigilo como exceção” (BRENOL, 2019, p. 82).

> “A LAI institui o direito previsto na Constituição de que todos têm a prerrogativa de receber dos órgãos públicos além de informações do seu interesse pessoal, também aquelas de interesse coletivo. Isto significa que a administração pública deve sistematizar a divulgação de suas ações e serviços, mas também deve estar preparada para receber demandas específicas.” (BRENOL, 2019, p. 82).

A LAI incentiva a cultura de acesso à informação no Brasil e estimulou jornalistas - que já vinham intensificando o uso de dados públicos desde a inclusão de computadores na redação e o acesso a internet - a buscarem e utilizarem ainda mais esse tipo de informação, uma vez que a lei resultou, também, na criação de canais legais de acesso e pedido. Assim, os dados podem ser disponibilizados espontânea e abertamente pelas instituições ou podem estar acessíveis mediante requisição.

> Esses dois mecanismos tornaram-se mais uma das formas pelas quais os jornalistas têm acesso a informações que constituem a apuração de suas pautas. Os dados abertos podem ainda possibilitar não apenas o acesso, mas também a leitura, análise, manipulação e cruzamento desses dados.” (BRENOL, 2019, p. 82)

Com o tempo, o atendimento às normas da LAI foi se aprimorando, o volume de dados acessíveis aumentou e mais jornalistas passaram a utilizar a LAI como instrumento de trabalho, para basear matérias jornalísticas e fiscalizar ações públicas, o que indica uma possibilidade de incorporação da lei na rotina de trabalho desses profissionais (BOTTREL, 2016). Nesse cenário, é possível que o planejamento das pautas passe a dar mais espaço para textos bem apurados e os jornalistas precisam ter habilidades para compreender e explicar os dados (GERALDES; SOUSA, 2016). Alguns veículos e instituições brasileiros contribuem para essa mudança ao difundir o uso da LAI e promover o políticas de dados abertos, como a Associação Brasileira de Jornalismo Investigativo (Abraji), a Open Knowledge Brasil, a Escola de Dados e a agência Fiquem Sabendo.

O acesso à LAI pode ser exercido por meio de transparência ativa, quando órgãos disponibilizam dados de forma espontânea, sem necessidade de solicitação,

e passiva, quando essas informações são fornecidas apenas quando solicitadas. O Portal de Transparência do Governo Federal no Brasil é um exemplo de transparência ativa e, junto com outras iniciativas, amplia o acesso aos dados públicos. Mas a leitura das bases de dados requer competências específicas (BRENOL, 2019).

Essas competências e técnicas de leituras de dados, próximas à prática científica, já eram usadas nas redações antes da chegada dos computadores (BRENOL, 2019). Uma das utilidades dessas técnicas é responder a críticas como a de não mostrar histórias relevantes, depender de assessorias, ser manipulável por grupos de interesse e de não comunicar de forma eficiente, frequentemente dirigidas aos jornalistas (MEYER, 2002). Assim, métodos científicos de pesquisa, como as enquetes, se popularizaram e seu uso incentivou a criação de institutos de pesquisa comandados pelas próprias empresas jornalísticas, a exemplo do DataFolha, do grupo Folha de São Paulo, o que permite a produção de dados e pautas próprios do veículo. Mesmo com esses recursos, há um conflito de tempos entre o jornalista e o instituto de pesquisa, pois para o jornalista há a urgência das informações factuais e do fechamento (BRENOL, 2019).

Com o tempo, repórteres também passaram a incorporar técnicas da Reportagem Assistida por Computador (RAC), começando por usar programas de edição, editoração e análise de informações, como buscas avançadas na internet, planilhas e estruturação de base de dados (BRENOL, 2019).

Nesse mesmo contexto, Meyer (2002) considerou importante que os jornalistas conhecessem o hardware do computador e tivessem habilidade para escolher softwares a serem usados como ferramentas de trabalho. Segundo o autor, os softwares se classificam como interativo, ou seja, aquele que responde a partir de comandos em um menu, e em “pacote”, ou batch mode, quando apresenta uma lista de comandos para receber como resultado o trabalho completo. Ainda, de acordo com Meyer, quanto mais fácil de aprender a usar o software, menos flexível ele é e, assim, jornalistas que aprendem a programar teriam mais controle e possibilidades se comparados com aqueles que apenas clicam em menus.

A ciência da computação associada ao jornalismo combina algoritmos, dados e conhecimento das ciências sociais, o que trabalha para aumentar a precisão no conteúdo produzido. Esse jornalismo computacional pode ser definido a partir das abordagens de reportagem assistida por computadores e uso de recursos das

ciências sociais, com o objetivo de explorar dados para pesquisar histórias. Essas ferramentas potencializam a fiscalização e investigação e envolvem a mineração de dados de interesse público (HAMILTON; TURNER, 2009).

Essa relação entre jornalismo e computador também pode ser entendida a partir da separação entre ciências da computação central, que diz respeito a operações matemáticas para o funcionamento da máquina, como a programação; e interativa, que lida com extração, modelagem e apresentações de informações a partir da interação homem-máquina, visualização gráfica e sistemas inteligentes (DIAKOPOULOS, 2012).

Para Stavelin (2013), o jornalismo computacional está relacionado aos seguintes critérios: estar focado em plataformas de discussão, pesquisa e narrativas de histórias; utilização de modelos computacionais para não só coletar, mas também analisar dados; e partir do pensamento computacional para desenvolver softwares voltados a soluções para o jornalismo.

Esse tipo de trabalho está ligado ao Jornalismo Digital em Base de Dados (JDBD), modelo em que os dados organizam e estruturam a produção, criação, apresentação e circulação jornalísticas. Esse tipo de jornalismo pode ser definido a partir de sete categorias: dinamicidade, automatização, inter-relacionamento/hiperlinkagem, flexibilidade, densidade informativa, diversidade temática e visualização (BARBOSA, 2007). Dessa forma, o conceito de JDBD é amplo e pode ser utilizado para estudar a relação das bases de dados com todas as dimensões do jornalismo, enquanto o Jornalismo Guiado por Dados (JGD) é “mais instrumental, metódico e técnico” (BRENOL, 2019). Assim, pode ser considerado que no JDBD os dados estão na estrutura que alicerça o jornalismo, não sendo enxergados somente como fontes de informação, como acontece no JGD, conceito que envolve técnicas de "RAC, visualização de dados, infografia, criação e manutenção de bases de dados e as políticas de acesso à informação e transparência pública de governos" (TRÄSEL, 2014, p. 106) organizadas nas etapas de coleta, limpeza, análise e visualização de dados (BRENOL, 2019).

Esse jornalismo tem como uma de suas principais características o uso de dados públicos que, ao serem construídos para atender a demandas sociais específicas, não são neutros. Utilizando essas bases como fontes principais, os jornalistas continuam a dar destaque às instituições que há muito tempo estão em evidência na mídia, como transmissoras de conhecimento por meio de declarações

e releases, permanecendo num oficialismo que é alvo de críticas há anos (TRÄSEL, 2014).

O que diferencia as bases de dados públicas de outras fontes oficiais é que, nos portais de transparência, os jornalistas têm acesso a relatórios completos e podem escolher os recortes e cruzamentos de dados que vão executar (BRENOL, 2019).

2.1.1 A visualização como ferramenta

Para conseguir, de fato, enxergar e interpretar dados, é preciso visualizá-los. Na definição usada por Aisch (2011), visualização de dados inclui até representações textuais de dados, como planilhas. Assim, para o autor, a questão não é se jornalistas de dados devem ou não utilizar visualizações, mas sim que tipo de visualização é melhor para cada situação, e quase sempre faz sentido ir além de uma visualização em planilhas. “Apenas tabelas definitivamente não são suficientes para nos dar uma visão geral de um conjunto de dados” (AISCH, 2011, p. 140, tradução da autora). O autor segue, afirmando que tabelas também não são o suficiente para identificar padrões nos dados.

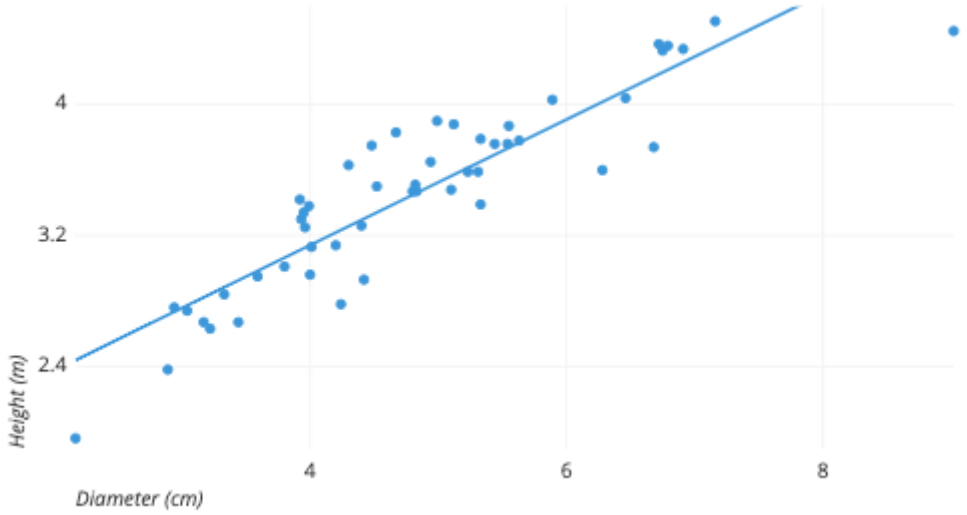
Não há garantia de que produzir algum tipo de visualização vai revelar uma história. O que o jornalista pode fazer para contar uma história a partir de dados é buscar por *insights* na visualização. Alguns desses *insights* podem já ser conhecidos, mas não provados, e outros podem ser novos e até surpreendentes. Ainda, alguns *insights* podem ser o começo de uma história enquanto outros podem ser o resultado de erros nos dados (AISCH, 2011).

Para encontrar esses *insights*, é preciso saber como visualizar os dados, o que pode ser feito de diferentes formas. Uma delas é a tabela que, de acordo com Aisch (2011), é poderosa quando se está lidando com uma quantidade pequena de dados, uma vez que ela mostra categorias e valores de uma forma estruturada e organizada e é ainda mais eficiente quando combinada com a habilidade de classificar e filtrar dados. Porém, é difícil comparar diferentes dimensões ao mesmo tempo em tabelas (no exemplo citado por Aisch, população por país ao longo do tempo).

Outro tipo de recurso é o gráfico, que permite o registro de dimensões dos dados em formas geométricas. Em um gráfico de dispersão, por exemplo, duas

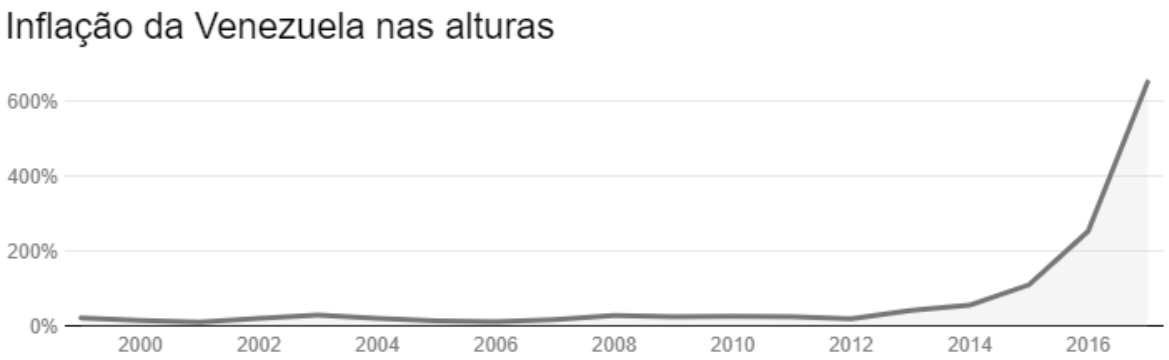
dimensões são mapeadas a partir das posições dos eixos x e y, com a possibilidade de mostrar uma terceira dimensão a partir da cor ou tamanho dos símbolos exibidos. Enquanto isso, gráficos de linha são adequados para evoluções temporais e gráficos de barras são bons para comparações entre dados categóricos. Também é possível colocar diferentes elementos uns sobre os outros, mostrando diversos ângulos do mesmo gráfico, para auxiliar na comparação entre pequenos grupos presentes nos dados; e usar diferentes escalas (por exemplo, linear e logarítmica) para explorar múltiplos aspectos dos dados observados (AISCH, 2011).

Gráfico 1 - Exemplo de gráfico de dispersão



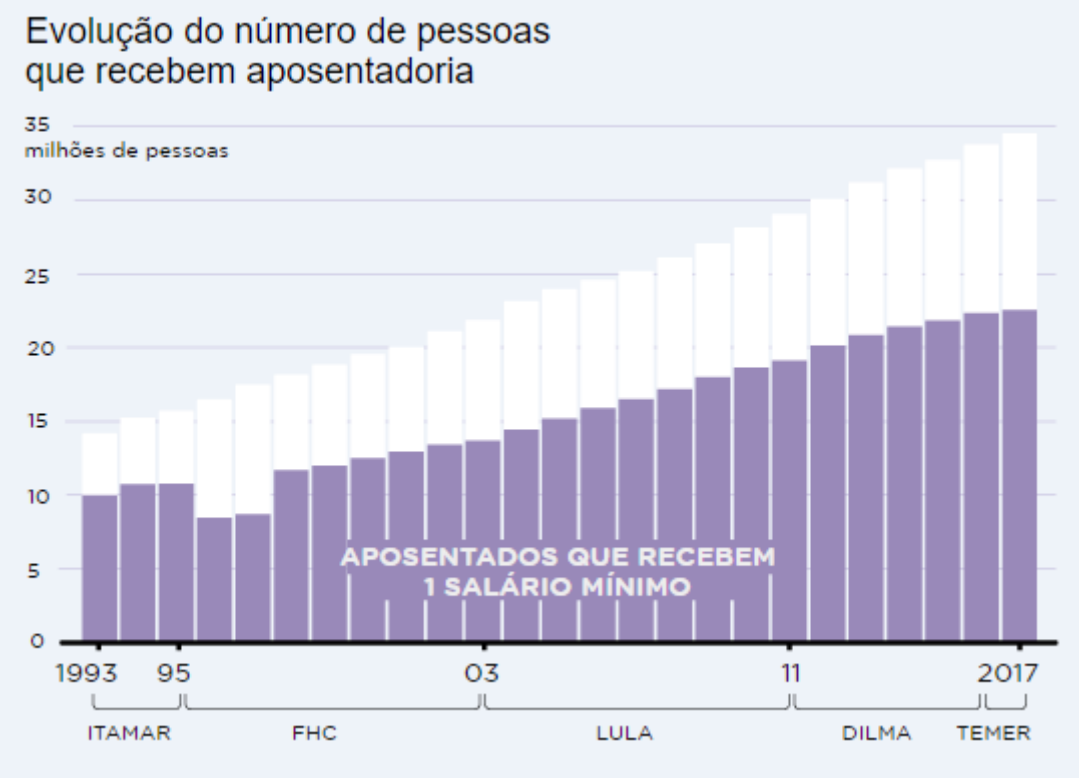
Fonte: A Complete Guide to Scatter Plots. Disponível em <https://chartio.com/learn/charts/what-is-a-scatter-plot/>. Acesso em: 16/04/2022.

Gráfico 2 - Gráfico de linha na reportagem Explorando o Arco Mineiro



Fonte: Explorando o Arco Mineiro. Disponível em <https://arcominero.infoamazonia.org/story>. Acesso em: 16/04/2022.

Gráfico 3 - Gráfico de barras em “A evolução do número de aposentados que recebem 1 salário mínimo”, do Nexo Jornal



Fonte: Nexo Jornal. Disponível em: <https://pp.nexojornal.com.br/Dados/2020/06/29/A-evolu%C3%A7%C3%A3o-do-n%C3%BAmero-de-aposentados-que-recebem-1-sal%C3%A1rio-m%C3%ADnimo>. Acesso em: 17/04/2022

Enquanto isso, mapas têm a capacidade de reconectar dados ao mundo físico e revelar padrões geográficos (AISCH, 2011).

Outro tipo importante de visualização é o grafo, elemento que mostra conexões entre peças de dados, representadas pelos nós. A posição desses nós é calculada por algoritmos que revelam estruturas em rede (AISCH, 2011).

Ao produzir uma visualização, o próximo passo é analisar e interpretar o que é mostrado. Para isso, Aisch sugere perguntas como “O que eu estou vendo nessa imagem? É o que eu esperava? É revelado algum padrão interessante? O que isso significa no contexto dos dados?” (AISCH, 2011, tradução da autora). Com esses questionamentos, é possível que o jornalista descubra que as belas visualizações produzidas acabaram não mostrando nenhuma informação relevante (AISCH, 2011).

Outra recomendação é que o jornalista documente seus caminhos e *insights*, para “lembrar para onde você foi, o que você viu lá e como você decidiu quais seriam os próximos passos” (AISCH, 2011, tradução da autora). Inclusive, o autor

recomenda documentar os objetivos e razões da decisão de iniciar a análise, no intuito de identificar vieses do próprio jornalista e evitar interpretações equivocadas.

Para observar em detalhes determinados padrões, também há meios de transformar os dados, como os citados por Aisch (2011): zoom (zooming) para observar detalhes; agregar (aggregation) para combinar diferentes dados num mesmo grupo; filtrar (filtering) para remover temporariamente pontos de dados que não são o seu foco; e remover valores aberrantes (outlier removal) para ignorar valores que não representam 99% do conjunto de dados (AISCH, 2011).

Sobre ferramentas, Aisch recomenda aquelas que podem ser usadas tanto para preparar quanto para visualizar os dados, uma vez que separar esses processos significa que os dados precisarão ser importados e exportados várias vezes. O autor faz uma relação de ferramentas para montar planilhas (LibreOffice, Excel e Google Docs); frameworks de programação estatística (R e Pandas); Sistemas de Informações Geográficas (Quantum GIS, ArcGIS e GRASS); bibliotecas de visualização (d3.js, Prefuse e Flare); preparação de dados, ou data wrangling (Google Rene e Datawrangler); e softwares de visualização (ManyEyes e Tableau Public) (AISCH, 2011).

3 PRECISÃO, OBJETIVIDADE E TRANSPARÊNCIA

A busca pela apreensão da realidade de forma a trazer informações que se aproximem da “verdade” é uma das tarefas com as quais o jornalista se ocupa constantemente. Ao mesmo tempo, essa realidade não pode ser apreendida em sua totalidade, uma vez que é possível acessar apenas fragmentos dela. Neste capítulo, explora-se como o Jornalismo Guiado por Dados (JGD) pode auxiliar numa compreensão mais ampla e objetiva da realidade.

3.1 O Jornalismo de Dados e a objetividade

A verdade, no sentido absoluto, filosófico, não pode ser atingida no processo jornalístico. Ao mesmo tempo, no sentido do que é revelado com precisão, ela pode e deve ser perseguida, na medida do que é possível investigar, apreender e entregar ao público no dia a dia (LISBOA e BENETTI, 2015). Dessa forma, a objetividade indica a correspondência entre o que é noticiado e o que aconteceu na realidade (GUERRA, 2008).

O que acontece é que as percepções, interpretações, pontos de vista e acesso às informações variam, fazendo com que a “verdade”, mesmo quando é retratada com observação à objetividade, possa ser mostrada de diferentes formas. Assim, não existe um jeito único de retratar a realidade e, ao mesmo tempo, para que o relato seja o mais preciso possível, é necessário que haja alguma forma de padronização. Esta padronização apenas pode ser atingida, disciplinadamente, a partir do método (LIPPMANN, 1920, p. 67). Esse método é uma das bases do Jornalismo de Dados. Este pode incluir hipóteses, oferece a possibilidade de maior transparência, quantifica fenômenos e cruza informações a partir de bases de dados (GEHRKE, 2017, p. 6).

O modelo de investigação utilizado pelo Jornalismo de Dados surgiu a partir da Reportagem Assistida por Computador (RAC), que teve desenvolvimento a partir dos anos 2000, com a popularização dos computadores. Nesse tipo de jornalismo, são utilizados dados públicos ou são criados repositórios utilizados como base para a produção jornalística (GEHRKE, 2017, p. 7). Com esse método mais objetivo, o Jornalismo de Dados tornou-se uma das principais respostas à crise econômica e identitária do jornalismo contemporâneo (TRÄSEL, 2014) e também uma forma de

jornalistas assumirem a responsabilidade de trabalharem em uma atividade semelhante à ciência, que, assim como o jornalismo, é uma resposta à necessidade de conhecimento e entendimento do ser humano (CRANBERG, 1989). Essa sede por conhecimento estava presente no jornalismo de precisão, reformulado para se tornar Reportagem Assistida por Computador (RAC) no final dos anos 1980 e início dos anos 1990 (CODDINGTON, 2014). O RAC trouxe a necessidade de que os jornalistas desenvolvessem habilidades com o computador (GEHRKE, 2017, p. 8), em especial porque, muitas vezes, notícias importantes são encontradas no meio de arquivos, bases de dados ou anotações (DADER, 2002) e a investigação dessas informações dispensa o uso da fala de fontes para confirmar acontecimentos (TRÄSEL, 2014).

Na superação desse jornalismo declaratório, a utilização de um método é importante. Para medir determinado fenômeno, é preciso definir critérios que vão nortear a análise, a partir de aspectos observáveis (MEYER, 2007, apud GEHRKE, 2017, p. 9). Em relação a isso, Guerra (2008) enumera três parâmetros para que a objetividade seja alcançada: a intenção do repórter em atingi-la, considerando apenas dados reais na apuração; o conhecimento do fato a partir da observação, entrevista e pesquisa documental rigorosas; e a redação criteriosa da notícia. Critérios como estes diferenciam o jornalismo e a ciência do senso comum, pois buscam conhecer a realidade a partir de um método (SPONHOLZ, 2009, p. 117, apud GEHRKE, 2017, p. 10). Ao mesmo tempo, a falta de recursos e a necessidade de publicações em ritmo acelerado exigem uma simplificação e limitam esse método (FRANCISCATO, 2006, apud GEHRKE, 2017, p. 10), mas não devem impedir que o jornalista siga determinados conceitos básicos da objetividade, que para Kovach e Rosenstiel (2004) são não acrescentar nada, não enganar o público, ser transparente sobre métodos e motivos, confiar apenas no próprio trabalho e ser humilde, aspectos cujo cumprimento também pode ser auxiliado pelo Jornalismo Guiado por Dados, que se torna cada vez mais importante nos ambientes jornalísticos, auxilia na descoberta de pautas inéditas e reforça a utilização do método científico (GEHRKE, 2017, p. 11). Inclusive, acredita-se que esse tipo de atividade ajudará o jornalismo a evoluir, fazendo mais do que decidir o que o público deve fazer, ao dar mais importância ao oferecimento de produtos confiáveis, com mais verificação e abertura sobre suas fontes e métodos (MEYER, 2004), num cenário marcado pela internet, que mudou o papel do jornalismo e exige que haja

um maior trabalho de reportagem (ANDERSON, BELL E SHIRKY, 2013), o que abre a possibilidade da utilização do Jornalismo de Dados como ferramenta para salvar o jornalismo a serviço do interesse público (TRÄSEL, 2014).

Assim, o Jornalismo de Dados tem como pré-requisito a utilização da objetividade como método e também modifica a noção desse conceito porque abre possibilidades de um jornalismo não mais sustentado por fontes declaratórias, em que o leitor pode verificar as informações fornecidas a qualquer momento (GEHRKE, 2017, p. 12).

Nesse cenário, a função do jornalista não é mais produzir informações iniciais, mas sim dar sentido ao amontoado de informações que chega ao público por meio da internet, combatendo a desinformação que também chega a partir desse meio (FONSECA, 2018).

3.2 A precisão e a verificação como elementos do jornalismo atual

Para Kovach e Rosenstiel (2004), a verificação separa o jornalismo do entretenimento, da propaganda, da literatura e da arte. Essa perspectiva, unida à transparência, está presente no conceito de jornalismo de precisão, apresentado por Meyer (1973) e presente no conceito do Jornalismo de Dados, que é a “aplicação da computação e dos saberes das ciências sociais na interpretação de dados, com o objetivo de ampliar a função da imprensa como defensora do interesse público” (TRÄSEL, 2014). Nesse jornalismo e no cenário que se apresenta com a internet, a verificação ganha um papel importante, pois ajuda a atravessar um ambiente de hiperconcorrência em que todos os discursos precisam entreter (FONSECA, 2018). No processo de verificação, as fontes documentais são indispensáveis e podem ser classificadas em três categorias, segundo Gehrke (2018a): arquivo documental (como leis, projetos e publicações); estatística (por exemplo, bases de dados públicas e séries históricas); e reprodução (conteúdo de colunas, blogs, sites e veículos jornalísticos).

O acesso a essas fontes é facilitado pelo ciberespaço, assim como às fontes declaratórias, que podem ser acessadas sem necessidade de deslocamento geográfico. Mas, para acessar todas as facilidades do ciberespaço, os jornalistas precisam dominar ferramentas e desenvolver habilidades técnicas para coleta e

análise de dados. Além de apontar para um novo mercado, essas questões abrem novas perspectivas para o ensino de jornalismo (FONSECA, 2018).

A habilidade de checagem de dados, vinculada ao Jornalismo Guiado por Dados, é uma dessas novas possibilidades a serem exploradas pelo ensino. Para Heravi (2017), jornalistas já formados precisam de treinamento específico em Jornalismo de Dados, em assuntos como coleta e limpeza de dados e análise estatística. Para estudantes, as primeiras habilidades a serem desenvolvidas seriam as de investigação jornalística.

Essas habilidades são úteis para atividades como a checagem de fatos, por exemplo, que pode verificar informações “que contenham números, comparações, dados estatísticos, que façam menção a documentos ou dados históricos” (FONSECA, 2018, p. 75). Para tais verificações, o jornalista precisa desenvolver habilidades para trabalhar com planilhas, fórmulas e gráficos.

Esse trabalho de checagem, por sua transparência no método, permite ao público julgar a validade das informações e identificar possíveis motivos do jornalista (KOVACH, ROSENSTIEL, 2004).

O Jornalismo de Dados, peça fundamental no processo de verificação de fatos, tem os dados como sua principal fonte de informação, o que exige a busca em bancos de dados, relatórios, contratos e outros documentos (FONSECA, 2018). É fundamental que essas habilidades sejam melhor desenvolvidas entre estudantes e jornalistas formados.

3.3 A transparência no Jornalismo de Dados

Numa reportagem publicada pelo jornal Folha de São Paulo sobre as chances de fraude na prova do Enem, “Estudo inédito indica alta chance de fraude em mil provas do Enem”, foi utilizado um modelo estatístico para a análise de dados públicos em busca de indícios de irregularidades, cujo método foi parcialmente divulgado (GEHRKE, 2018b). Em seu artigo “Transparência no método como valor para o jornalismo”, Marília Gehrke analisa a reportagem com base nos parâmetros clareza nas fontes; abertura de pesquisas, testes e análises; e correção de erros e atualização (GEHRKE, 2018b).

Esses parâmetros têm capacidade de expor diferentes aspectos da transparência, que é importante para que o leitor acredite que o jornalismo diz a

verdade, que está justificada no discurso jornalístico (LISBOA, BENETTI, 2015). Assim, a verdade e a verificação são essenciais para o jornalismo, como mencionado por Kovach e Rosenstiel (2004), que diziam que a primeira obrigação do jornalismo é com a verdade, sua primeira lealdade é com os cidadãos, sua essência é a disciplina da verificação, seu produto deve ser independente de quem é coberto e do poder, que ele deve abrir espaço para crítica e precisa apresentar o que é significativo de forma interessante e relevante. Considerando esses aspectos, a transparência pode ser interpretada como uma abertura no processo de reportagem, principalmente em relação a "fontes e métodos, escolhas, passo a passo adotado e suas limitações" (GEHRKE, 2018b, p. 4). De acordo com Karlsson (2010), a transparência tem duas vertentes, uma referente à abertura e divulgação do processo de produção do conteúdo, que, segundo ele, pode ser feita a partir da divulgação de links, e outra que tem a ver com a participação dos usuários, a partir da interatividade, para que o público possa monitorar, checar e intervir no processo que, de acordo com Ward (2011), é caracterizado pela busca dos jornalistas pela verdade e pelo relato verdadeiro. Esta busca diferencia o jornalismo de outras fontes de informação menos confiáveis, que, segundo Philip Meyer (em entrevista), devem ser combatidas tornando-se a verdade visível. Nesse processo pode-se entender que a transparência é fundamental. E esta, para Weinberger (2009), pode ser obtida por meio da divulgação de links. Além de incluir essa possibilidade, o Jornalismo de Dados tem uma marca de colaboração, que contribui para a transparência.

Nesse tipo de jornalismo, informações colhidas de bases de dados são a matéria prima, além de "leis, memorandos de atas, estudos científicos, pesquisas, relatórios e publicações em sites de redes sociais" (GEHRKE, 2018b, p. 6). Como, na maioria das vezes, esses dados são públicos, o leitor, que hoje em dia já está acostumado a lidar com números, fatos e estatísticas, pode refazer o caminho do jornalista e confirmar a veracidade das informações, principalmente quando há hiperlinks no texto (MIELNICZUK, 2003) e o jornalista explica com detalhes o caminho que fez.

Voltando à descrição da reportagem "Estudo inédito indica alta chance de fraude em mil provas do Enem", para ela foi desenvolvido um modelo estatístico que analisou 3 milhões de gabaritos de provas realizadas entre 2011 e 2016, a partir de microdados do Inep, considerando para análise os alunos que estavam entre os 10% com melhores notas. Além disso, os repórteres também fizeram investigação

tradicional, *in loco*, para identificar os suspeitos de fraude. A conclusão foi de que houve fraude em pelo menos 1250 provas do Enem, que tinham um padrão de respostas parecido, o que torna improvável que não tenha ocorrido alguma infração (GEHRKE, 2018b). A reportagem também conta com infográficos para melhor apreensão das informações. O modelo é mais rígido que os utilizados para encontrar fraudes em outras provas (MARIANI, 2018), mas não foi divulgado pela Folha, o que mostra um problema em relação à transparência, com divulgação de poucas informações para que o leitor faça sua própria análise. Por exemplo, ao explicar como chegou a nomes de pessoas suspeitas de fraude, a reportagem diz somente que cruzou dados de ingressantes nas universidades com informações do Enem.

Um dos parâmetros para estimular a transparência nos jornalistas, segundo Gehrke (2018b), é a abertura de pesquisas, testes e análises, para que o leitor possa entender procedimentos técnicos, escolhas e variáveis. Esse aspecto é especialmente importante no Jornalismo Guiado por Dados. “O entendimento completo que pode gerar a verificação pelo público e também pelas fontes requer a liberação, na íntegra, dos procedimentos adotados” (GEHRKE, 2018b, p. 12). Isso pode incluir a divulgação do código de programação utilizado, contribuindo, no processo, para a colaboração no jornalismo. Ao mesmo tempo, parâmetros de transparência não devem ser aplicados somente ao Jornalismo Guiado por Dados, mas também em “qualquer procedimento que envolva uma apuração jornalística criteriosa” (GEHRKE, 2018b, p. 13).

3.3.1 A cultura hacker

A transparência é um dos valores da ética *hacker*, que está intrinsecamente ligada à liberdade de informação e, como Träsel (2018) observou, não está presente apenas em profissionais de tecnologia: esses valores éticos e cultura *hacker* também podem ser encontrados entre jornalistas especialistas em dados. Ao estudar a equipe de Jornalismo de Dados do jornal O Estado de São Paulo, o autor identificou comportamentos como a adoção do valor de transparência da ética *hacker*, interesse na manipulação e entendimento de artefatos técnicos e o incentivo ao uso da criatividade para lidar cada vez melhor com dados na redação.

O compartilhamento de código-fonte com o público, também característico da cultura *hacker*, acontece na redação e desafia a ideia tradicional de competitividade

entre empresas jornalísticas. Outras diferenças observadas são a rotina e a liberdade de instalar programas (uma vez que os membros do time de Jornalismo de Dados têm seus próprios computadores). Esses comportamentos podem ser resultado, também, de aspectos em comum entre jornalistas e *hackers*, como a tendência à curiosidade, amor pela tecnologia e cooperatividade (TRÄSEL, 2018).

Todas essas novas dinâmicas surgem em um momento em que, mais do que nunca, a audiência impacta no formato da notícia (NUNES, 2018). A ascensão do Jornalismo de Dados e a inclusão de programadores nas redações podem ser vistas como uma resposta à crise financeira das empresas jornalísticas, numa tentativa de atrair o público (TRÄSEL, 2018). Essas práticas foram iniciadas por veículos nativos digitais e posteriormente incorporadas por grandes empresas, onde a presença de novos especialistas tem potencial, também, para uma mudança estrutural, pois esses profissionais têm domínio técnico sobre as linguagens e softwares que hoje perpassam, quase inevitavelmente, todo o trabalho jornalístico, influenciando tanto os processos quanto os produtos (NUNES, 2018).

3.3.1.1 Software de código aberto

O código aberto (*open source*) pode ser definido como um desejo prático por programação aberta e uma crença filosófica na responsabilidade social por meio da liberdade e abertura (COLEMAN, 2004). Ele é caracterizado por uma transferência de conhecimento, sem relações de compra e venda ou contratuais, que consiste no compartilhamento de informações relevantes com um conjunto indefinido de outros atores, sem qualquer recompensa imediata e com o propósito de contribuir para o desenvolvimento em conjunto (LEWIS; USHER, 2013). Assim, os projetos de código aberto são motivados mais pelo bem comum e interesse da comunidade que por lucro e propriedade (TURNER, 2005). A ideia básica por trás disso é que o software funciona melhor quando está livremente disponível e é programado em conjunto (LEWIS; USHER, 2013).

“Em projetos de código aberto, usuários podem acessar, modificar e distribuir livremente o código-fonte, assim, permitindo que uma ideia bem sucedida se desenvolva rápido, já que ela é copiada para outro lugar e reconstruída por outros” (LEWIS; USHER, 2013, p. 607). Assim, a cultura de código aberto, que tem na

transparência um de seus valores, pode ser entendida como uma parte da cultura *hacker* e está interligada com o desenvolvimento da internet (KELTY, 2008).

Conectando essa cultura ao jornalismo, é importante lembrar que a transparência é um dos ideais do jornalismo e o jornalismo digital aumenta as possibilidades dela ser alcançada (PHILLIPS, 2010). Dessa forma, com a transparência do código aberto, a visão a respeito do jornalismo pode ser modificada, passando-se a enxergar as notícias não mais como um produto final, mas sim como um conjunto de interações mais fluido, ao qual os usuários poderiam contribuir (ROBINSON, 2011) .

4 FERRAMENTAS PARA O JORNALISMO DE DADOS

Muitas vezes apenas observar um conjunto de dados e organizá-lo de forma manual não é o suficiente ou necessita de um trabalho exaustivo para que sejam encontrados padrões ou informações relevantes para o material jornalístico. Lidar com esses dados de forma automatizada, utilizando ferramentas de programação, pode tornar o trabalho mais simples e revelar informações que dificilmente seriam notadas à primeira vista. Em grande parte das vezes, pode-se chegar ao resultado desejado realizando-se operações que lidam com todo o conjunto de dados de uma só vez, como, por exemplo, utilizar essas ferramentas para cruzar informações do IBGE e dos nomes de logradouros brasileiros para encontrar quantos são nomes femininos e quantos são masculinos.

Por outro lado, em operações em que o volume de dados é maior, o que exige a separação desse conjunto de dados em partes menores e realização de operações separadas em cada uma delas, para que a máquina seja capaz de processá-las, são utilizadas ferramentas de Big Data. Esse termo pode ser definido como um conjunto de técnicas usadas para extrair informações compreensíveis e úteis de um grande conjunto de dados, com valores complexos e desconhecidos. A expressão foi documentada pela primeira vez em um artigo da NASA de 1997. O Big Data está relacionado a uma enorme velocidade, variedade e volume de dados que podem ser estruturados ou desestruturados (GULIA; RATRA, 2019).

Utilizar essas técnicas para acessar e interpretar a infinidade de informações disponíveis na rede pode originar um formato diferente de jornalismo, em que todas as etapas do processo atravessam a internet. Nesse contexto, a computação abre novas possibilidades benéficas, como maior transparência nos órgãos públicos, novas descobertas experimentais e inovação nos modelos de negócios. Além disso, a utilização dessas mesmas ferramentas contribui para a mudança na composição das redações, com a aproximação dos cientistas da computação junto aos profissionais que desenvolvem as diferentes funções que compõem a produção do Jornalismo de Dados e procuram ir além da notícia (GOMES JR., 2014). Esses profissionais podem se utilizar do Big Data em situações caracterizadas por três critérios (os três V’s do Big Data): grande Volume de dados, grande Velocidade possível (capacidade necessária para processar esses dados) e Variedade (os

dados são obtidos através de diferentes fontes). A isso pode ser adicionado o Valor (é importante que os dados acrescentem algo a seu utilizador) (GOMES JR., 2014).

4.1 Convergência e Jornalismo de Dados nas redações

Observando as mudanças e possibilidades trazidas pela computação e grande volume de informações disponíveis, algumas redações mudaram suas estruturas para acessar diferentes mídias e formas de trabalho. Mesmo com essa busca de empresas pela convergência, com uso de recursos como as bases de dados, muitas vezes há dificuldades na integração e adaptação de ferramentas, profissionais e métodos (BARBOSA; ALBAN, 2013).

> A convergência jornalística é um processo multidimensional que, facilitado pela implantação generalizada das tecnologias de telecomunicação, afeta âmbito tecnológico, empresarial, profissional e editorial dos meios de comunicação, proporcionando uma integração e ferramentas, espaços, métodos de trabalho e linguagens anteriormente separados, de forma que os jornalistas elaboram conteúdos que se distribuem através de múltiplas plataformas, mediante as linguagens próprias de cada um. (SALAVERRÍA; GARCÍA AVILÉS E MASIP, 2010, p. 59, tradução da autora).

No processo de aplicação da convergência nas empresas jornalísticas, as bases de dados são agentes fundamentais. Nesse sentido, existe o Jornalismo Digital em Base de Dados (JDBD), que usa essa ferramenta em sua estrutura, organização, composição e apresentação dos conteúdos (BARBOSA, 2007). As tecnologias digitais facilitam esse processo e podem estar presentes nas áreas empresarial, profissional e editorial do veículo (BARBOSA; ALBAN, 2013).

Suzana Barbosa e Renato Alban fizeram uma análise do jornal Correio e do site Correio24horas. Nessa análise, eles descobriram que o uso de base de dados no jornal ainda era iniciante. São usadas plataformas principalmente para a convergência de conteúdo, estruturando textos, fotos, infográficos, dados, vídeos e áudios em bases de dados. Mas, enquanto há possibilidades de uso de bases de dados para tornar o conteúdo da versão on-line diferente da versão impressa e fazer complementações de conteúdo na versão web, esse não parecia ser o caminho que estava sendo adotado pelo jornal (BARBOSA; ALBAN, 2013).

As redações do site e do impresso usavam bases de dados diferentes e separadas e não há alterações de linguagem e complementação de conteúdo quando passa-se do impresso para o on-line (BARBOSA; ALBAN, 2013).

Outras redações também seguiram esse caminho e começaram a implementar a convergência, como é o caso de O Estado de São Paulo, Folha de São Paulo e O Globo, que também avançaram em iniciativas de Jornalismo de Dados. “Quando a informação era escassa, a maior parte de nossos esforços estavam voltados a caçar e reunir dados. Agora que a informação é abundante, processá-los tornou-se mais importante” (FLEW et al., 2012). O Jornalismo de Dados, ao exigir uma maior qualificação profissional, inicia uma discussão sobre o papel da imprensa na democracia, ao mudar o tipo e qualidade do conteúdo ofertado (MANCINI; VASCONCELLOS, 2016).

Esse tipo de jornalismo pode ser definido pela associação da utilização de dados, investigação jornalística e utilização de tecnologias que permitem uma apuração e visualização mais complexa dessas informações (BRADSHAW, 2014). Nele, são feitos a produção, tratamento e cruzamento de dados, tornando a recuperação da informação, a apuração da reportagem, a circulação em diferentes plataformas e a geração de visualizações mais eficientes (TRÄSEL, 2013).

Ao mesmo tempo, a prática do Jornalismo de Dados ainda está muito associada com o uso de infográficos e há questionamentos sobre se aqueles jornalistas capazes de garimpar e analisar dados, mas que não fazem infográficos, são mesmo jornalistas de dados (MANCINI; VASCONCELLOS, 2016).

A cobertura investigativa depende de materiais gerados ou reunidos pelo próprio repórter, para isso, o jornalista utiliza computadores, na reportagem com auxílio de computador (RAC), e o Jornalismo de Dados seria uma vertente do RAC, porque envolveria uso de computadores, conhecimento de estatística, sistemas computacionais e métodos das ciências sociais (Flew et al., 2012; Hamilton and Turner, 2009; Gray et al., 2014; Hackett, 2013; Howard, 2014 apud MANCINI; VASCONCELLOS, 2016). O RAC une técnicas das ciências sociais ao jornalismo, enquanto o jornalismo de dados e o jornalismo computacional também se relacionam, além das ciências sociais, com a cultura de dados abertos, assim, o jornalismo de dados estaria mais próximo da cultura dos dados abertos, junto com o trabalho computacional (MANCINI; VASCONCELLOS, 2016). A representação de fenômenos é possível a partir da quantificação deles, assim, enquanto a

quantificação transforma os fenômenos em dados, a análise os transforma em conhecimento (STRAY, 2014).

Quanto à produção de infográficos, muitas vezes os únicos lembrados são aqueles mais criativos e esteticamente atraentes, mas é importante observar que aqueles infográficos menos esteticamente elaborados e mais fáceis de produzir, mas com valor analítico, também devem ser lembrados (MANCINI; VASCONCELLOS, 2016). Para Mancini e Vasconcellos (2016), há uma diferença entre reportagem com dados e reportagem de dados. Segundo eles, a primeira se apropria de dados de forma ilustrativa e a segunda tem nos dados o fundamento da pauta e, na relação entre eles, a condução da história.

Na dimensão interpretativa do Jornalismo de Dados, a reportagem, além de apresentar o conteúdo e contexto, deve trazer uma análise da relação entre os dados. Nesse sentido, o jornalista também pode trazer suas próprias análises ou análises de entrevistados (MANCINI; VASCONCELLOS, 2016).

No aspecto comunicativo, considera-se os meios de visualização de dados e como eles são utilizados para promover a compreensão da reportagem. O estudo desses e outros aspectos facilita a compreensão da escola que vai do Jornalismo com Dados para o Jornalismo de Dados. Para compreender os dados, jornalistas podem usar comparações estatísticas para descobrir causas, consequências ou implicações, aplicando competências de extração, estruturação, análise e visualização de dados (MANCINI; VASCONCELLOS, 2016).

A relação entre os dados e o jornalismo se modificou no século XXI. Um dos meios que ilustra essa modificação é o blog Estadão Dados, onde eram publicadas visualizações e análises por jornalistas programadores. Hoje desativado, ele foi criado para disseminar a prática do Jornalismo de Dados na redação. Porém, de acordo com Del Vecchio-Lima e Specht (2021), o blog poderia explorar matérias mais humanizadas.

O Estadão Dados usava algoritmos para organizar e tornar dados compreensíveis, para que a audiência faça suas próprias análises. Algumas das seções do blog eram o Gráfico do Dia, com publicações sobre temas variados; Permanentes, com séries estatísticas atualizadas constantemente; e Séries, material com temática específica, que apresentava explicações e análises (DEL VECCHIO-LIMA; SPECHT, 2021). Para analisar esse blog, Del Vecchio-Lima e Specht usaram critérios de transparência (informações sobre a metodologia utilizada

para o tratamento e reelaboração de dados), estruturação do conteúdo (adaptabilidade a outros meios de veiculação e existência de hiperlinks), interatividade (personalização de visualização, compartilhamento e comentários), contextualização/análise (feitos pelos próprios jornalistas ou por fontes) e humanização (presença de histórias ilustrativas).

O primeiro conteúdo analisado do blog foi “São Paulo, uma cidade dos anos 70”, com um mapa que apresentava a idade média das construções da cidade. Junto ao material há um pequeno texto com uma análise de jornalistas e, na reportagem completa, aparecem personagens e fontes especializadas (humanização). O outro conteúdo analisado, “Veja o desempenho dos candidatos na média dos institutos”, tem quatro gráficos sobre pesquisas de candidatos às eleições de 2014, sem links para reportagens relacionadas, mas com mais interação dos leitores, a partir de comentários, três deles com respostas do jornalista que produziu o conteúdo (DEL VECCHIO-LIMA; SPECHT, 2021).

O estudo de Del Vecchio-Lima e Specht (2021) pretende ressaltar a importância da utilização da tecnologia para uma melhor contextualização, mais transparência, oferecimento de possibilidades de investigações diversas do próprio leitor, além do conteúdo apresentado na reportagem nos materiais jornalísticos e humanização no material jornalístico.

4.2 Usos e possibilidades da programação

O jornalista que utiliza ferramentas mais avançadas da computação, no senso comum, é alguém mais qualificado em termos técnicos e também mais sobrecarregado, uma vez que esse trabalho poderia ser feito por uma equipe multidisciplinar. O uso dessas ferramentas está ligado à cultura e cognição dos indivíduos, além da disponibilização do instrumental técnico (DEL VECCHIO-LIMA; SPECHT, 2021). De acordo com Stray (2014), o Jornalismo de Dados tem quatro estágios: quantificação, processo que envolve humanos e máquinas em que é preciso transformar a realidade em dados quantificáveis; análise, estágio que exige conhecimentos técnicos específicos, em que dados são transformados em conhecimento e em que o jornalismo se aproxima da ciência, baseando-se em matemática, estatística e lógica; comunicação, processo mais eficiente quando conectado a pessoas e histórias, que leva os dados extraídos à audiência; e ação,

estágio em que, a partir da comunicação, esses dados se tornam agentes transformadores.

As bases de dados são fundamentais nesse processo. Elas são consideradas por alguns pesquisadores como uma forma de cultura simbólica, observando-se sua ubiquidade, e podem ser analisadas a partir de diferentes estratégias (DEL VECCHIO-LIMA; SPECHT, 2021).

Uma das funções do Big Data é a mineração de dados, ou data mining, processo que consiste em extrair informações previamente desconhecidas e compreensíveis, encontrando padrões úteis, em um grande volume de dados (RAVAL, 2012). Para esse tipo de trabalho, Gulia e Ratra (2019) elencam as ferramentas árvore de decisão (*classification trees*), regressão logística (*logistic regression*), redes neurais (*neural networks*), técnicas de clusterização (*clustering techniques*), mineração de regras de associação (*association rule mining)* e aprendizado de máquina (*machine learning*). A seguir, cada uma delas é brevemente descrita.

Árvore de decisão: ferramenta utilizada para classificar uma variável com número limitado de valores possíveis, atribuindo-se cada unidade a uma categoria com base em propriedades qualitativas. O resultado é uma estrutura em formato de árvore com nós que representam associações e grupos. Essa ferramenta pode ser usada quando a tarefa envolve previsões ou classificação de resultados (GULIA; RATRA, 2019).

Regressão logística: técnica estatística que permite a previsão de valores a partir de uma série de observações. Ela fornece um método usado para prever a probabilidade de ocorrência das variáveis independentes em uma função (GULIA; RATRA, 2019).

Redes neurais: algoritmos inspirados pelo sistema nervoso de animais que formam uma rede que consiste em camadas de *input* (entrada), camadas escondidas e camadas de *output* (saída). Nesse modelo, diferentes tipos de dados são colocados nas camadas de input e, a partir de uma técnica de tentativa e erro, algoritmos são usados para ajustar os pesos até que ele atenda a critérios de travamento pré-determinados (GULIA; RATRA, 2019).

Técnicas de clusterização: usadas para agrupar dados em subclasses, ou clusters. Uma delas é a *Knearest neighbor*, que encontra a lacuna entre o registro

real e os dados de treinamento. Depois disso, ela atribui cada registro à classe que está mais próxima em todo o conjunto de dados (GULIA; RATRA, 2019).

Mineração de regras de associação: essa técnica envolve o uso de aprendizado de máquina para buscar padrões em uma base de dados identificando associações de if then[2] frequentes. O antecedente (if) e a consequência (then) são duas partes da regra de associação. O primeiro é descoberto a partir dos dados e o segundo é um item descoberto com o antecedente (GULIA; RATRA, 2019).

Aprendizado de máquina: softwares com a capacidade de aprender a partir dos dados, com o objetivo principal de fazer previsões baseadas em propriedades conhecidas. Gulia e Ratra (2019) citam os exemplos de uso dessa técnica para encontrar a diferença entre mensagens que são ou não são spams e aprender as preferências de usuários para fazer recomendações (GULIA; RATRA, 2019).

Já para a análise de dados, as autoras elencam diferentes ferramentas, que podem ser classificadas de acordo com o estágio dos dados em que atuam, sendo divididas em ferramentas de coleta, que auxiliam na formação de um conjunto de dados, ao ajudar na coleta de informação em múltiplos locais e/ou de acordo com padrões pré-determinados, encontrando palavras-chave e fazendo análises subjetivas para reunir um grupo específico de dados que se quer analisar; de armazenamento e *frameworks*[3], que têm a função de acomodar os dados e servir de base para futuros trabalhos, como certas ferramentas do Apache; de extração e filtragem, úteis para obter dados específicos da internet e estruturá-los, sendo aquelas para raspagem de dados (*scraping*), que consiste em minerar mecanicamente informações da internet, algumas das mais utilizadas; e de limpeza e validação, estágio em que a necessidade e relevância dos dados encontrados são verificadas e a retirada de informações não úteis diminui o tempo de processamento (GULIA; RATRA, 2019).

---

[2] Função que permite selecionar quais comandos serão executados dependendo de uma condição (UC SANTA CRUZ, 2016).

[3] Estrutura com software de função genérica que pode ser modificada de acordo com o que é programado nela (RIEHLE, 2000)

5 METODOLOGIA E ANÁLISE

> *"O nosso conteúdo é estruturado no trabalho conjunto entre jornalismo e dados abertos, pois acreditamos que este é um dos caminhos mais efetivos para a construção de uma democracia transparente e responsável."*
> *(AGÊNCIA TATU, 2017)*

Ter uma cultura de dados abertos requer mais esforço. Nas dez peças analisadas nesta monografia, houve comprometimento em divulgar o método e procedimentos utilizados, seja no tratamento, interpretação ou visualização dos dados, em conta do jornalista ou do veículo na plataforma GitHub. Isso evidencia uma preocupação a mais com a transparência e contribuição para a sociedade.

O GitHub tem a função de registrar *scripts* utilizados em qualquer tipo de projeto que utilize linguagens de programação. O que não seria diferente com peças jornalísticas. O conteúdo analisado foi escolhido com base em uma pesquisa em busca de veículos com destaque por terem investido no Jornalismo de Dados. Foram verificadas listas em sites especializados em coberturas jornalísticas e, após encontrados os nomes de veículos e jornalistas que desenvolvem esse tipo de trabalho no Brasil, verificou-se quais deles divulgavam, na plataforma GitHub, a metodologia e código-fonte utilizados, com a seleção de uma peça jornalística de cada um dos dez encontrados, utilizando os critérios de originalidade, representatividade e diversidade, pois, conforme Machado e Palacios (2018), quanto mais original a peça, mais adequada para os propósitos de pesquisa. Além disso, quanto mais representativa de uma tendência maior as chances de ser incluída e quanto mais destoante da tendência mais chance de servir como contraprova .

Considerando o conteúdo analisado, a metodologia escolhida foi o modelo de pesquisa de estudo de caso por ilustração desenvolvido durante dez anos pelo Grupo de Pesquisa em Jornalismo On-line (GJOL), do Programa de Pós-Graduação em Comunicação e Cultura Contemporâneas da Universidade Federal da Bahia (MACHADO; PALACIOS, 2018). Utilizando este método, primeiramente foi realizada uma análise preliminar das peças jornalísticas, para um mapeamento inicial, com organização em duas planilhas: uma com perguntas, para descrição mais detalhada do conteúdo (APÊNDICE A), e uma mais objetiva, para organizar as peças em categorias (APÊNDICE B).

Posteriormente, foram estabelecidas, a partir de particularidades do *corpus* observadas na etapa de análise preliminar deste trabalho, as categorias de análise *transparência, profundidade, abertura, relevância* e *interatividade*, considerando-se as características do material divulgado no GitHub, a forma de utilização das linguagens de programação, sua adequação aos objetivos da peça jornalística e seu produto, considerando o nível de necessidade de utilização desses processos ao trazer à luz informações que não seriam descobertas de outras formas.

Apresentar os veículos e materiais escolhidos e analisar o material para responder o problema de pesquisa são os objetivos deste capítulo.

5.1 Peças analisadas

As peças analisadas foram escolhidas a partir de sua disponibilidade na conta do GitHub do veículo ou dos jornalistas. Foram escolhidas peças sobre temas diversos, com algumas que abordam assuntos similares, para comparação de diferentes abordagens a respeito de um mesmo assunto. A seguir estão as peças escolhidas.

Conheça os nomes mais populares de 2021 em cada estado brasileiro[4], da Agência Tatu: reportagem publicada em 22 de dezembro de 2021 que analisa dados do Portal Oficial do Registro Civil para verificar quais foram os nomes mais comuns no Brasil em 2021. Descobriu-se que Miguel foi o nome mais registrado em nove estados. Há um mapa interativo com os nomes mais comuns em cada estado brasileiro: para observar o nome do estado, nome mais escolhido e número de crianças nascidas registradas com esse nome, basta passar o cursor sobre o estado no mapa, que também mostra os nomes mais escolhidos por meio de distinção de cores. O texto que acompanha o mapa descreve os resultados com atenção a padrões e singularidades e traz, ainda, a explicação de um antropólogo sobre o porquê dos mesmos nomes se repetirem.

Fiquem Sabendo lança aplicativo para consulta de remuneração de pensionistas[5], da Fiquem Sabendo: anúncio, publicado no site da Fiquem Sabendo

---

[4] Disponível em <https://www.agenciatatu.com.br/noticia/conheca-os-nomes-mais-populares-de-2021-em-cada-estado-brasileiro/>. Acesso: 31/03/2022

[5] Disponível em <https://fiquemsabendo.com.br/transparencia/app-dados-pensionistas/>. Acesso: 31/03/2022.

em 26 de julho de 2021, do aplicativo criado pela agência para consultar remunerações de pensionistas a partir de dados disponibilizados pela Controladoria-Geral da União depois de denúncias da Fiquem Sabendo. Desenvolvido em parceria da agência com o cientista de dados Fernando Barbalho, o aplicativo tem o intuito de facilitar e democratizar o acesso a essas informações, com funcionalidades de cruzamento de dados, visualização de série histórica, pesquisa pelo nome do servidor instituidor, aplicação de filtros e download de pílulas dos dados.

Questões do Enem na mira de Bolsonaro são eficientes em testar conhecimento[6], da Folha de São Paulo: reportagem feita a partir de análises estatísticas para avaliar se as questões sobre temas polêmicos entre conservadores - e com veto cogitado por Bolsonaro - têm capacidade técnica para avaliar o conhecimento dos participantes. Publicada em 20 de novembro de 2021, a reportagem analisa, a partir de microdados do Inep, 24 questões do Enem criticadas por políticos conservadores ou que fazem menção à ditadura militar, considerando quatro critérios de avaliação de qualidade utilizados na literatura científica da área. A conclusão foi que as perguntas atendem aos critérios. A reportagem buscou posicionamento do Inep, que não respondeu. A peça também traça um histórico dos ataques ao Enem feitos por políticos e complementa o conteúdo com uma análise para verificar se candidatos provavelmente transexuais (aqueles que utilizavam nome social) acertaram mais uma questão específica, sobre um dialeto usado por essa parcela da população, criticada por Bolsonaro por supostamente favorecer esses candidatos. As fontes buscadas pela reportagem foram um especialista em avaliação de questões e um ex-presidente do Inep, que falam sobre o método de avaliação utilizado e os aspectos políticos envolvidos.

90% da nota do Enem é influenciada por fatores econômicos e culturais, indica análise feita por GaúchaZH com uso de inteligência artificial[7], de GZH: análise publicada em 26 de julho de 2019 que buscou verificar se há relação entre a condição socioeconômica de candidatos gaúchos e seus desempenhos no Enem.

---

[6] Disponível em <https://www1.folha.uol.com.br/educacao/2021/11/questoes-do-enem-na-mira-de-bolsonaro-sao-eficientes-em-testar-conhecimento.shtml>. Acesso: 31/03/2022.

[7] Disponível em <https://gauchazh.clicrbs.com.br/educacao-e-emprego/noticia/2019/07/90-da-nota-do-enem-e-influenciada-por-fatores-economicos-e-culturais-indica-analise-feita-por-gauchazh-com-uso-de-inteligencia-artificial-cjyk4u1jq054u01msvn171rj2.html>. Acesso em: 31/03/2022.

Depois de cálculo estatístico com auxílio de inteligência artificial, a reportagem constatou que 89,9% da nota na prova é influenciada por esses fatores e identificou as cinco condições que mais influenciaram nas notas. A reportagem também contém visualizações com as médias de notas obtidas em alguns estados e as mil maiores notas por renda familiar e por presença de computador em casa, além de explicações e exemplos sobre a relação entre esforço, condições socioeconômicas e nota, com uma fonte especializada e depoimentos de uma aluna de Psicologia moradora do bairro Lomba do Pinheiro. Ao final, a reportagem também traz explicações sobre os procedimentos adotados nas análises.

Rua: substantivo (ainda) masculino[8], de Gênero e Número: reportagem em vídeo publicada em 15 de março de 2017 que constatou, a partir do cruzamento de um banco de dados do IBGE Nomes com uma relação de todos os nomes de logradouros brasileiros, que apenas 20% destes têm nomes femininos. No vídeo, que é composto de uma série de ilustrações, são mostradas as características da nomeação dos logradouros brasileiros ao longo da história e, mais especificamente, em alguns estados.

Explorando o Arco Mineiro[9], de InfoAmazonia: reportagem multimídia com depoimentos de mineiros, empresas, acadêmicos, indígenas, políticos e ativistas sobre a disputa por áreas de mineração na Venezuela. O conteúdo inclui texto, vídeos, fotos e gráficos numa interface criada especialmente para a reportagem, que explora a questão detalhadamente utilizando, para complementar a narrativa, visualizações de dados sobre o Preço global do petróleo bruto ao longo dos anos e casos confirmados de malária na Venezuela, com informações extraídas de bases de dados internacionais.

A evolução do número de aposentados que recebem 1 salário mínimo[10], do Nexo Jornal: análises e visualizações de dados baseadas no artigo “O caráter social do salário mínimo: da política de valorização à reposição da inflação”[11] a respeito dos valores recebidos em aposentadorias no Brasil. O material foi publicado

[8] Disponível em <https://www.generonumero.media/rua-substantivo-ainda-masculino/>. Acesso em: 31/03/2022.
[9] Disponível em <https://arcominero.infoamazonia.org/?lang=pt>. Acesso em: 31/03/2022.
[10] Disponível em <https://pp.nexojornal.com.br/Dados/2020/06/29/A-evolu%C3%A7%C3%A3o-do-n%C3%BAmero-de-aposentados-que-recebem-1-sal%C3%A1rio-m%C3%ADnimo>. Acesso em: 31/03/2022.
[11] Disponível em <https://pp.nexojornal.com.br/opiniao/2020/O-car%C3%A1ter-social-do-sal%C3%A1rio-m%C3%ADnimo-da-pol%C3%ADtica-de-valoriza%C3%A7%C3%A3o-%C3%A0-reposi%C3%A7%C3%A3o-da-infla%C3%A7%C3%A3o>. Acesso em: 31/03/2022.

em 19 de junho de 2020 e contém cinco representações gráficas sobre o número de pessoas que recebiam salário mínimo de aposentadoria em 2017 comparado ao total de aposentados; a evolução temporal do número de pessoas que recebem aposentadoria; a evolução temporal da porcentagem dos aposentados que recebem um salário mínimo; o salário mínimo ao longo do tempo; e a valorização real do salário mínimo.

O que 15 mil tweets revelam sobre seu candidato[12], de O Estado de São Paulo: análise de 15.654 posts no Twitter de pré-candidatos à presidência em 2018 publicada em 9 de maio de 2018. O objetivo foi identificar as peculiaridades de cada candidato, buscando as palavras que cada um deles citava muito e outros citavam pouco. A reportagem tem elementos interativos e ensina, no início, como ler os gráficos apresentados. Depois, ela exibe o gráfico das postagens de cada candidato acompanhado de um breve comentário. Ao fim, a reportagem descreve os procedimentos utilizados.

Coronavírus avança para o interior do Brasil; veja evolução em mapa[13], do UOL: publicada em 16 de abril de 2020, a peça mostra a evolução dos casos de covid-19 pelo Brasil a partir de uma linha do tempo com informações em um mapa. No texto, a reportagem faz um histórico e descrição da situação brasileira em relação ao vírus até o momento de publicação.

Atlas da notícia[14], parceria do Instituto para o Desenvolvimento do Jornalismo (Projor) com o Volt Data Lab: mapeamento de veículos produtores de notícias brasileiros com primeira versão publicada em 2017. A metodologia utiliza tanto pesquisa própria quanto contribuições de terceiros. Na página, além do mapeamento completo, há análises com abordagem de questões como a formação de desertos de notícias.

---

[12] Disponível em <https://infograficos.estadao.com.br/politica/eleicoes/2018/o-que-15-mil-tweets-revelam-sobre-seu-candidato/>. Acesso em: 31/03/2022.
[13] Disponível em <https://noticias.uol.com.br/saude/ultimas-noticias/redacao/2020/04/16/coronavirus-avanca-para-o-interior-do-brasil-veja-evolucao-em-mapa.htm> Acesso em: 31/03/2022.
[14] Disponível em <https://www.atlas.jor.br/dados/app/>. Acesso em: 01/04/2022

5.2 Páginas no GitHub

O GitHub é um ambiente de transparência em software que incorpora características de redes sociais e possibilita uma atividade interdisciplinar, numa união entre jornalismo e tecnologia (TSAY; DABBISH; HERBSLEB, 2014). Assim, a plataforma é um meio para que jornalistas produzam conteúdo com código aberto e abre a possibilidade de interação com tecnólogos para construção de um jornalismo com novos valores, sem deixar de lado os princípios fundamentais da prática, tornando-a, assim, mais relevante e participativa na cultura digital (LEWIS; USHER, 2013).

O desenvolvimento de software não costuma ser um processo linear. À medida que se escreve o código, são repetidos procedimentos de testes, descoberta de erros e correções desses erros. Se algo dá errado no meio do caminho, procurar o problema em cada linha de código pode ser difícil e lento, enquanto a possibilidade de restaurar e comparar com versões anteriores torna esse processo mais simples (DEVMOUNTAIN, |20-?|). A necessidade de encontrar um meio eficiente de controlar e acessar diferentes versões de um código-fonte foi o que motivou a criação do Git, em 2005, por Linus Torvalds, também criador do kernel, componente que liga o *hardware* às operações executadas no sistema operacional Linux. O Git é um sistema de controle de versão, ou seja, ele tem a função de gravar e permitir o gerenciamento de alterações feitas no código-fonte. Gratuito, de código aberto e contando com um sistema de ramificações do *script*, ele também permite que grandes times de programadores trabalhem de forma independente em suas próprias versões do código (PERKEL, 2018).

Para funcionar, o Git precisa ser instalado no computador. Por sua vez, o GitHub opera totalmente na nuvem, funciona como uma base de dados on-line que hospeda o Git e possibilita o compartilhamento do código-fonte a partir de fora da máquina onde ele foi originalmente criado, além de disponibilizar uma interface gráfica acessível, novas ferramentas de controle e gerenciamento e oferecer a possibilidade de acesso remoto ao repositório por pessoas autorizadas pelo criador original do conteúdo (DEVMOUNTAIN, |20-?|).

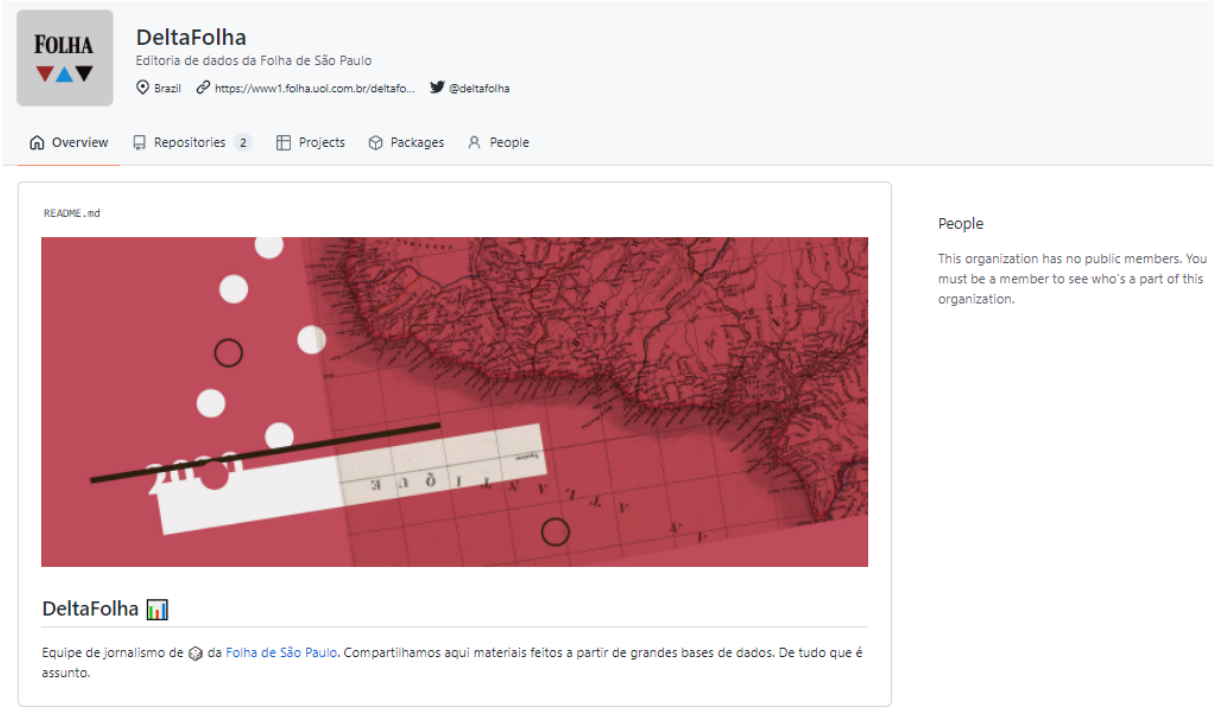
Ao contrário do Git, o GitHub é uma empresa com fins lucrativos. Ela foi criada em 2008 por Chris Wanstrath, J. Hyett, Tom Preston-Werner e Scott Chacon e comprada pela Microsoft em 2018. Hoje, a plataforma conta com 73 milhões de

usuários e mais de 200 milhões de repositórios (GITHUB, 2022). Essencialmente, o GitHub tem a função de ajudar pessoas a resolver problemas a partir da cooperação no desenvolvimento de software. Qualquer pessoa, em todo o mundo, pode acessar o código-fonte compartilhado no GitHub, sugerir mudanças e melhorias e tirar dúvidas. Os desenvolvedores de um projeto também podem trabalhar em equipe por meio da plataforma, delegando tarefas, discutindo questões e criando ramificações do código-fonte para trabalhar em modificações de forma segura, sem alterar o funcionamento do código original. Essas versões do código podem ser alteradas por outros membros da equipe e depois incorporadas ao projeto principal. Além disso, uma vez que o código está aberto, outras pessoas podem utilizá-lo em seus próprios projetos (GITHUB, 2016). Do ponto de vista da cooperação que o GitHub possibilita, sua forma de operação pode ser comparada à da Wikipédia, plataforma em que pessoas de diferentes partes do mundo constroem e revisam peças de informação de forma colaborativa.

Assim, a publicação do código-fonte do material jornalístico no GitHub se encaixa nas duas vertentes da transparência sugeridas por Karlsson (2010): ela possibilita tanto a abertura e divulgação dos processos de produção e métodos utilizados quanto a participação do público, que pode não só avaliar por si mesmo a validade dos processos, como também sugerir mudanças e aperfeiçoamentos no código por meio do GitHub.

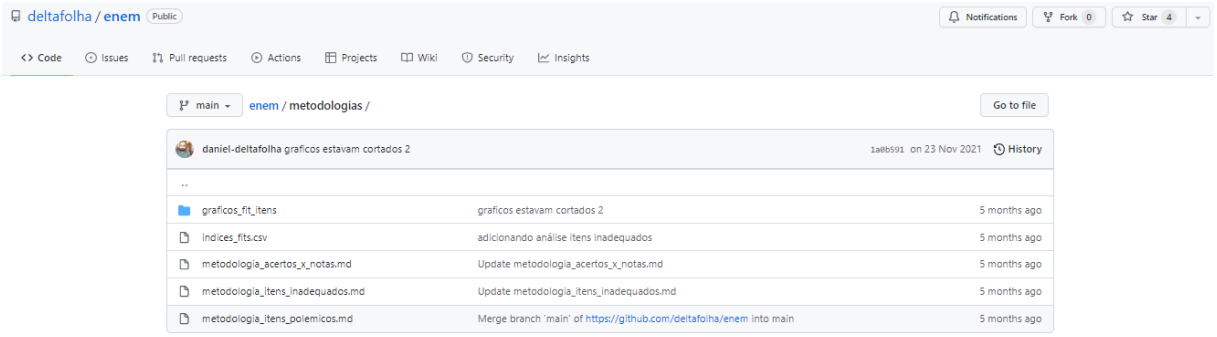
Nesta monografia, as contas do GitHub avaliadas pertencem aos próprios veículos ou a jornalistas que tiveram a iniciativa de compartilhar o código-fonte utilizado. Cada um escolheu expor os dados de uma forma diferente, dentro das possibilidades da plataforma.

Captura de tela 1 - Página inicial da DeltaFolha no GitHub



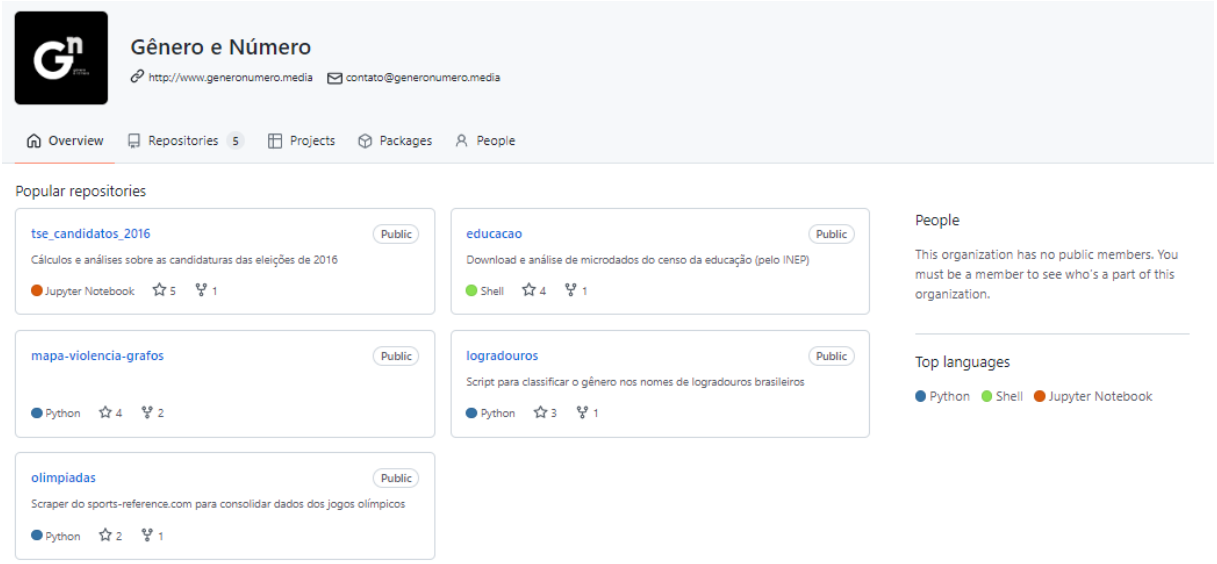
Fonte: github.com

Captura de tela 2 - Repositório no GitHub sobre as metodologias utilizadas em reportagens que analisaram microdados do Enem



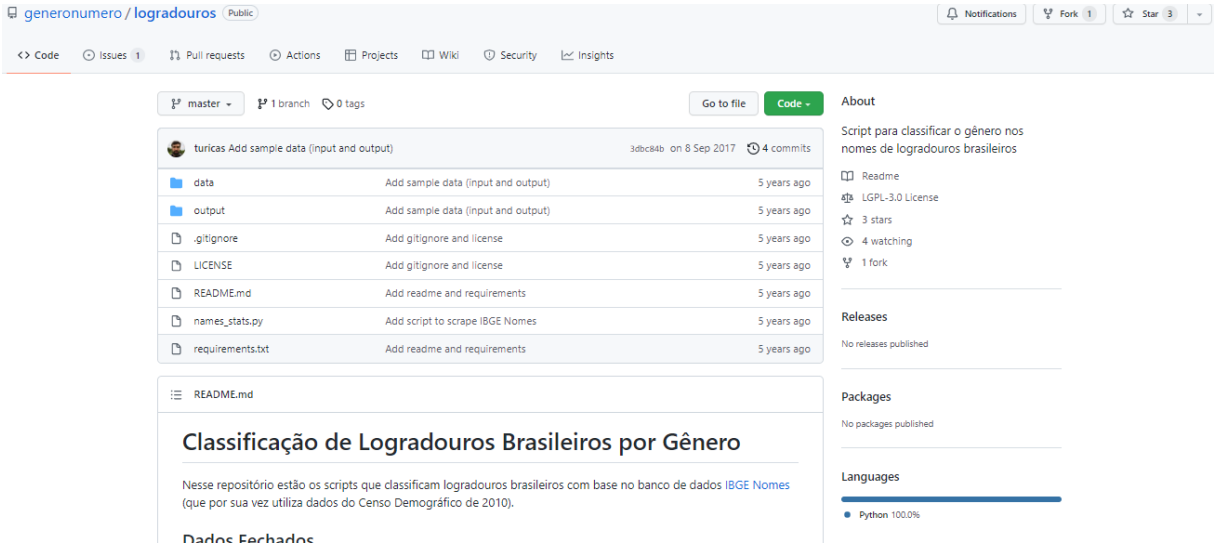
Fonte: github.com

Captura de tela 3 - Página inicial da Gênero e Número no GitHub



Fonte: github.com

Captura de tela 4 - Repositório no GitHub sobre “Rua: substantivo (ainda) masculino”, da Gênero e Número



Fonte: github.com

5.3 Método: passo a passo

O método adotado neste trabalho foi o modelo de estudo de caso por ilustração desenvolvido pelo Grupo de Pesquisa em Jornalismo On-line (GJOL), do Programa de Pós-Graduação em Comunicação e Cultura Contemporâneas da Universidade Federal da Bahia. O método foi criado num contexto de tentativas de mapeamento de modificações do jornalismo no cenário do ciberespaço e se propõe

a apresentar conceitos-chave desse novo jornalismo, identificando características de produções jornalísticas e possibilitando o entendimento tanto dos fundamentos teóricos quanto das particularidades das organizações (MACHADO; PALACIOS, 2018).

Esse modelo de análise foi escolhido por ter sido aprimorado especificamente para o jornalismo digital, que é foco neste trabalho. Machado e Palacios descrevem:

> "Neste modelo híbrido, procedimentos de pesquisa qualitativa e quantitativa são ações complementares no processo contínuo de compreensão conceitual sobre a produção de informações nas organizações jornalísticas no ciberespaço nas sociedades contemporâneas." (MACHADO; PALACIOS, 2018).

Os objetos de pesquisa deste trabalho são as peças de Jornalismo de Dados e Jornalismo com Dados selecionadas e suas respectivas descrições e *script*s disponíveis na plataforma GitHub. A análise busca especificamente identificar como se caracterizam os processos de divulgação dos métodos no GitHub; que linguagens de programação e abordagens foram utilizadas; e que tipo de resultado elas produzem, aliadas aos outros elementos presentes no material publicado. O foco do trabalho não será o conjunto de competências dos jornalistas envolvidos ou análise dos veículos citados. O foco também não está em análises de recepção do público ou alcance do conteúdo, mas sim na descrição do material postado no GitHub, processos utilizados e resultados finais.

Para definir o material a ser estudado, primeiramente foram buscadas iniciativas emergentes de Jornalismo de Dados no Brasil, em listas de referências como a elaborada pelo Dados Abertos Pernambuco[15]. Posteriormente, verificou-se, dentre as iniciativas encontradas, quais possuíam contas no GitHub com informações sobre o material publicado, chegando-se a dez contas, ao todo. Em determinados casos, o veículo não divulga as informações em seu nome, mas sim jornalistas envolvidos o fazem. Então, uma peça jornalística foi selecionada em cada uma das contas, considerando-se os critérios de originalidade, representatividade e diversidade, para adequação aos propósitos de pesquisa, inclusão em tendências e contraprovas, respectivamente. Assim, o *corpus* é composto de dez peças

[15] Disponível em <https://www.dadosabertospernambuco.com.br/jornalismodedadosbr>. Acesso em: 01/04/2022.

jornalísticas em diferentes formatos, com diferentes propostas. Foram considerados, inclusive, materiais com objetivo principal de servir como fonte para outras peças jornalísticas. No quadro a seguir são descritos os veículos com material analisado.

Quadro 1 - Veículos com trabalhos analisados nesta monografia

Nome	Descrição	Criação	Site	GitHub
Agência Tatu	Startup criada por três estudantes de Jornalismo da Universidade Federal de Alagoas (Ufal), a agência nasceu como um veículo laboratorial com foco em Jornalismo de Dados.	2017, Alagoas	agenciatatu.com.br	github.com/lucasthaynan
Fiquem Sabendo	Agência independente de dados públicos especializada na Lei de Acesso à Informação. O objetivo da Fiquem Sabendo é divulgar informações de interesse público que não são divulgadas pelas instituições governamentais, tendo feito mais de 600 publicações com dados inéditos em seu website. A agência também possui a newsletter Don't LAI to me e um canal no YouTube, além de ter sido usada como fonte para mais de 2 mil reportagens publicadas pela imprensa.	2015, São Paulo	fiquemsabendo.com.br	github.com/FiquemSabendo
Folha de São Paulo	O jornal paulista, de maior circulação no Brasil, tornou-se pioneiro em Jornalismo de Dados em 2012, quando criou, em parceria com o programa Knight, do Centro Internacional para Jornalistas (ICFJ), o FolhaSP Dados (REFERÊNCIAS NO BRASIL, 2020). Posteriormente, a seção passou a se chamar DeltaFolha e tem atuação constante no veículo, em especial no Twitter.	1921, São Paulo	folha.uol.com.br	github.com/deltafolha
Gaúcha ZH (GZH)	Jornal digital que reúne conteúdos produzidos pelo jornal Zero Hora e pela rádio Gaúcha, também com conteúdo exclusivamente digital. Tem 92 mil assinantes (GAÚCHA ZH, 2020). GZH não tem uma equipe exclusiva para jornalismo de dados, mas tem repórteres com competências para produzir esse tipo de pauta.	2017, Rio Grande do Sul	gauchazh.clicrbs.com.br	github.com/leonardojs1981
Gênero e Número	Empresa social focada em jornalismo de dados para análise de questões relacionadas a gênero e raça, com objetivo de qualificar debates pela equidade. "A partir de linguagem gráfica, conteúdo audiovisual, pesquisas, relatórios e reportagens multimídia alcançamos e informamos uma audiência interessada no assunto" (GÊNERO E NÚMERO, |20--?|).	2016, Rio de Janeiro	generonumero.media	github.com/generonumero
InfoAmazonia	Veículo independente que utiliza dados, mapas e reportagens geolocalizadas para contar histórias sobre a Amazônia. O portal cruza notícias com dados para "melhorar a percepção sobre os desafios para a conservação da floresta" (INFOAMAZÔNIA, |20--?|).	2012, São Paulo	infoamazonia.org	github.com/InfoAmazonia

Nexo Jornal	Jornal digital que tem como um de seus objetivos de criação ampliar o acesso a dados e estatísticas. O jornal possui, entre seus princípios editoriais, o de transparência, para construir uma relação de confiança e interação com sua audiência (NEXO JORNAL, 2022).	2015, São Paulo	nexojornal.com.br	github.com/Nexo-Dados
O Estado de São Paulo	Em 2012, o jornal paulista criou o Estadão Dados, time dedicado à análise e visualização de dados (TRÄSEL, 2018). Hoje, a seção de dados do jornal não está mais no blog “estadaodados.com”, mas sim em “estadao.com.br/infograficos”. Além disso, a conta no Twitter do Estadão Dados não tem novas atividades desde 2019.	1875, São Paulo	estadao.com.br	github.com/estadao
UOL	Empresa de conteúdo, serviços e produtos da internet do Grupo UOL PagSeguro. Publica notícias sobre os mais diversos assuntos e reportagens especiais (UOL, 2021).	1996, São Paulo	uol.com.br	github.com/juditecyprreste
Volt Data Lab	Agência independente de dados abertos responsável, junto ao Instituto do para o Desenvolvimento do Jornalismo (Projor), pelo Atlas da Notícia, portal que mapeia veículos jornalísticos no Brasil (ATLAS DA NOTÍCIA, 2022).	2014, São Paulo	voltdata.info	github.com/voltdatalab

Fonte: a autora (2022)

A metodologia também sugere a elaboração de um Protocolo de Estudo de Caso, um roteiro para a pesquisa, com definição de períodos, técnicas de coleta de dados, procedimentos e condutas a serem adotadas, organizados nas seções visão global do projeto, procedimento de campo, determinação das questões e guia para a elaboração do relatório.

No entanto, nem todas as seções se mostraram necessárias no caso desta análise. Assim, foram incluídas somente as seções de procedimento e determinação das questões. Foram elaboradas quatro questões com a finalidade de explorar o objeto empírico de forma objetiva. A intenção da análise foi, justamente, observar como foram divulgados a metodologia e o código-fonte na plataforma GitHub, quais foram as linguagens de programação utilizadas, quais os tipos de abordagem utilizados e que resultados foram observados, em conjunto com os outros elementos presentes no material publicado.

Dessa forma, o primeiro passo foi a leitura do material publicado nos respectivos sites dos veículos, na íntegra, para uma maior familiaridade com o assunto e questões envolvidas. Posteriormente, houve análise dos arquivos postados no GitHub a fim de responder às seguintes questões pré-estabelecidas: 1.

Como se caracteriza a descrição da metologia utilizada e a disponibilização do código-fonte?; 1.1. É possível compreender todos os processos a partir da descrição textual do conteúdo no GitHub?; 1.2. O código-fonte é divulgado de maneira completa, possibilitando a reprodução do conteúdo por inteiro apenas a partir das informações presentes?; 2. Quais foram as linguagens de programação utilizadas?; 2.1. Que características dessas linguagens as tornam adequadas para o desenvolvimento do projeto em específico?; 3. Quais foram as abordagens utilizadas?; 3.1. Que resultados esses tipos de abordagem produziram no conteúdo final?; 4. Que camadas de sentido os elementos escolhidos para complementar as análises e visualizações de dados adicionaram ao material final?.

De acordo com Machado e Palacios (2018), a terceira parte do método é a definição conceitual, em que, a partir da bibliografia e do material analisado, são escolhidas categorias de análise, que devem estar fundamentadas no referencial teórico, dividem o material em seus elementos e componentes ao mesmo tempo que se relacionam com o todo e sintetizam aspectos relevantes e contraditórios. Assim, foram criadas, a partir da base teórica e particularidades do objeto de pesquisa observadas na etapa de análise preliminar deste trabalho, cinco categorias: *transparência, abertura, interatividade, relevância e profundidade*. Elas foram desenvolvidas especificamente para esta pesquisa.

5.4 Objeto de pesquisa

Visto que o material analisado é composto, em uma parte, de informações divulgadas na plataforma GitHub e, em outra, em veículos de notícias, os dois foram separados para análise de cada um de acordo com suas características específicas. O *corpus* inclui a descrição textual do método no GitHub, arquivos com o código-fonte do material, visualizações de dados, texto e eventualmente vídeos e fotos. A seguir são descritos os critérios de classificação escolhidos para melhor definição do material estudado.

Foram entendidos como peças jornalísticas qualquer conteúdo com o objetivo de trazer informações relevantes de forma mais acessível ao público, podendo ser uma reportagem, com elementos como texto, imagens, vídeos e áudios junto às análises e visualizações de dados, apenas visualizações ou tabelas, contanto que tenham função de selecionar e processar dados com objetivo de tornar informações

mais acessíveis. Também foram identificados os bancos de dados dos quais foram extraídas as informações para análise, que foram classificados entre privados, públicos ou de organizações internacionais. Os bancos de dados públicos identificados foram aqueles provenientes do Portal Oficial de Registro Civil, Ministério da Defesa, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), Instituto Nacional do Seguro Social (INSS), Fundação Oswaldo Cruz (Fiocruz) e Secretaria Especial de Comunicação Social (Secom). Enquanto isso, os bancos de dados privados foram os provenientes da Empresa de Correios e Telégrafos (ECT) e Twitter e as organizações internacionais foram o Fundo Monetário Internacional (FMI), Organização dos Países Exportadores de Petróleo (OPEC) e Organização Mundial da Saúde (OMS).

Entre as outras fontes documentais e especializadas, oficiais e não oficiais, foram observados documentos em texto, professores, dirigentes de organizações e pessoas “comuns”.

As linguagens de programação encontradas foram HTML, JavaScript, Python, R e SQL. Criado em 1990, o HTML (Hyper Text Markup Language) é utilizado para estruturação de páginas na web e sua praticidade na criação de *home pages*, permitindo incluir conteúdo multimídia, como vídeos, imagens e áudios, foi um dos fatores para a popularização da internet (PINHEIRO, 1997). Também projetado para a web, o JavaScript foi desenvolvido pelos criadores da linguagem Java a partir de uma necessidade de mais dinamismo nos websites (RAUSCHMAYER, 2014), sendo usada na criação de páginas interativas que podem ser acessadas em uma variedade de dispositivos. Por sua vez, Python é simples e clara, o que torna o seu aprendizado mais fácil para pessoas com pouco conhecimento em programação, mas não deixa de ser multifuncional, com capacidade de administrar e grandes projetos e de ser usada em áreas como inteligência artificial, banco de dados, biotecnologia, animação 3D e aplicativos móveis (MENEZES, 2010). Já o R é uma linguagem que tem o propósito de resolver problemas gráficos e estatísticos, com funções especiais de exibir, organizar e armazenar dados (POVOA; MANZIONE; WENDLAND, 2011). Por fim, a SQL (Structured Query Language) é uma das linguagens mais utilizadas para comandos em bancos de dados relacionais, possibilitando criação de tabelas, campos, índices, atribuição de permissões a usuários e consulta de dados (BATISTA; FILHO; PIMENTEL; MARTINS, 2019).

Além das linguagens, foram identificadas as abordagens escolhidas para processar e organizar os dados em cada caso. Nesse processo, as abordagens encontradas foram raspagem de dados, árvore de decisão, limpeza, cálculo de frequência, tokenização, concatenação e outros métodos de processamento de dados. A raspagem de dados, ou, no termo de origem em inglês, *web scraping,* é uma técnica que permite extrair dados de diferentes sites na web para uma única planilha ou base de dados para tornar sua visualização e análise mais fáceis (SIRISURIYA, 2015). Enquanto isso, a árvore de decisão é composta por algoritmos de aprendizado de máquina (*machine learning*) que nomeiam casos de treino representados por uma tupla[16] de valores de atributos e uma identificação de classe (SU; ZHANG, 2006). Já a limpeza de dados consiste em consertar ou remover dados que não contribuem ou atrapalham o processo de análise, que podem estar incorretos, com erro de formatação, duplicados ou incompletos (CHU; ILYAS; KRISHNAN, 2016). Ao mesmo tempo, o cálculo de frequência mede, a partir de algoritmos, quantas vezes um resultado ocorre em determinado intervalo. E a tokenização é o processo de quebrar uma linha de texto em frases, palavras, símbolos ou outros elementos, chamados de tokens (GAUR; RENU; VERMA, 2014). Por fim, a concatenação é a junção de duas cadeias de caracteres em uma (EXCRIPT, |20--?|).

A temática de cada uma das peças também foi verificada, classificando-os entre cotidiano, política, educação, igualdade, direitos humanos, meio ambiente, economia, política, saúde e informação.

5.4.1 Publicações no GitHub

Um dos objetivos deste trabalho é analisar como se caracterizam as descrições de métodos e procedimentos e a disponibilização do código-fonte na plataforma GitHub. Assim, a primeira análise feita foi em relação à descrição do conteúdo disponibilizado em arquivos, ou seja, a introdução ao material. O primeiro passo nessa análise foi a contagem de caracteres (com espaços) e palavras. A média do número de caracteres foi 2.354,2 enquanto a média do número de palavras foi de 371,8. Porém, houve variação considerável: o texto mais curto, a

[16] Estrutura semelhante a uma lista, utilizada em Python

respeito do Atlas da Notícia, tinha apenas sete caracteres, ao mesmo tempo que o mais longo, sobre o aplicativo de pensões da Fiquem Sabendo, tinha 8.111.

A estrutura de todos os textos está organizada em um título e subtópicos, além dos textos, com exceção de duas, que têm apenas título, sem subtópicos. Quatro dos dez textos de explicação têm dois subtópicos principais, mas o número varia entre zero e sete. A seguir, é feita uma análise descritiva de cada um dos textos, considerando o critério de transparência de abertura de pesquisas, testes e análises (GEHRKE, 2018b), avaliando se, com as informações contidas no texto descritivo do material publicado no GitHub, o leitor é capaz de entender os métodos e procedimentos adotados. Os tópicos de análise escolhidos foram nomeação dos procedimentos desenvolvidos (1), descrição das razões de escolha desses processos (2), descrição do conteúdo nos arquivos disponibilizados (3), indicação da base de dados utilizada como fonte (4), descrição do passo-a-passo do processo (5), menção às bibliotecas e ferramentas utilizadas (6) e utilização de hiperlinks para complementar o conteúdo (7).

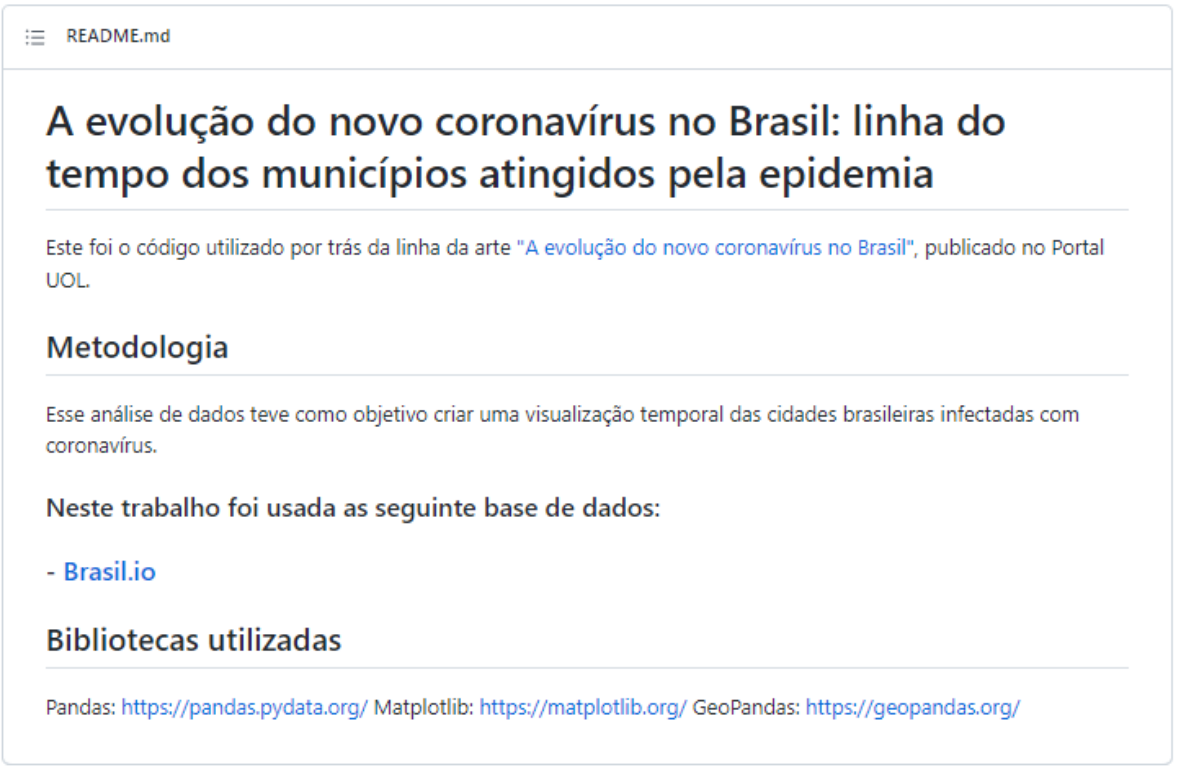
Quadro 2 - Tópicos atendidos por cada material disponibilizado no GitHub

Nome	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Nomes mais registrados no Brasil em 2021	S	S	S	S	N	S	S
Metodologia	S	S	S	S	S	S	S
Material e métodos da reportagem "Questões do Enem na mira de Bolsonaro são eficientes em testar conhecimento"	S	N	N	S	S	S	S
Analise_enem_2	S	N	S	N	N	S	N
Classificação de Logradouros Brasileiros por Gênero	S	N	N	S	S	S	S
Digging into the Mining Arc	N	N	N	N	S	S	S
01. Salário Mínimo	N	N	S	S	S	N	S
O que 15 mil	S	N	S	S	S	S	S

tweets revelam sobre o seu candidato							
A evolução do novo coronavírus no Brasil: linha do tempo dos municípios atingidos pela epidemia	N	N	S	S	N	S	S
An-site	N	N	N	N	N	N	N

Fonte: a autora (2022)

Captura de tela 5 - Descrição do repositório no GitHub para “Coronavírus avança para o interior do Brasil; veja evolução em mapa”



Fonte: github.com

Nomes mais registrados no Brasil em 2021[17], sobre a reportagem “Conheça os nomes mais populares de 2021 em cada estado brasileiro”: o texto indica que a ferramenta desenvolvida para tratamento dos dados foi um *web scraper* que extraiu os 50 nomes mais registrados de cada estado e município em 2021 do Portal Oficial do Registro Civil (com hiperlink para o site). O texto então prossegue

[17] Disponível em <https://github.com/lucasthaynan/ranking_nomes_br>. Acesso em: 02/04/2022.

com a descrição do conteúdo gerado como resultado, disponível em dois arquivos no formato .csv, e finaliza com a descrição do banco de dados utilizado e da razão da utilização do método escolhido.

Metodologia[18], sobre o aplicativo descrito em “Fiquem Sabendo lança aplicativo para consulta de remuneração de pensionistas”: no texto, é dito que a equipe fez leituras, pré-processamentos, extraiu, elaborou uma versão sintética, separou e cruzou dados. Ao descrever os procedimentos, a equipe menciona que a coluna de identificação do pensionista no Portal de Transparência é utilizada como chave comum entre as tabelas de cadastro e remuneração, para garantir a correspondência mesmo em casos de pensionistas homônimos e que, para lidar com mais de um registro de um mesmo pensionista, foi adicionada uma coluna que indica a existência de várias pensões, com remuneração do pensionista indicada em cada uma das linhas, nesses casos. O texto descreve, ainda, que o arquivo *graphs_on_demand.Rmd* contém instruções para operar a interface gráfica, que os dados resumidos do Portal da Transparência estão na tabela *consolidated_data.csv*, que comentários em *query_consolidated_data.sql* reproduzem a consulta para extração dos dados resumidos do banco de dados, que instruções para visualizações intermediárias estão em *query_join_tables.sql* e que a consulta para extração dos dados individualizados do último mês disponível está em *query_microdata.sql.* A base de dados indicada como fonte é o Portal de Transparência e o trabalho desenvolvido para cada funcionalidade também é descrito no texto, que menciona a utilização do Rmarkdown e de ferramentas como o shinnyapps.io. A descrição também traz hiperlinks em todas as suas seções. Em sua estrutura, o texto primeiramente introduz o repositório, descreve a base de dados utilizada, depois explica a história da divulgação dos dados dos pensionistas, reivindicada pelo próprio Fiquem Sabendo, detalha a estrutura do repositório, as transformações aplicadas nos dados. Ainda, o texto destaca pontos de atenção, que o leitor deve observar ao ler e interpretar as informações, informações de contato e instruções sobre créditos ao utilizar os dados divulgados.

Material e métodos da reportagem "Questões do Enem na mira de Bolsonaro são eficientes em testar conhecimento"[19]: o texto especifica que foi

[18] Disponível em <https://github.com/FiquemSabendo/pensionistas>. Acesso em: 02/04/2022.
[19] Disponível em <https://github.com/deltafolha/enem/blob/main/metodologias/metodologia_itens_polemicos.md>. Acesso em: 03/04/2022.

feito um modelo estatístico baseado em três parâmetros. Estes estão de acordo com o método TRI (Teoria de Resposta ao Item), utilizado pelo Inep para corrigir e dar notas. Neste caso não há repositório de arquivos, com os elementos complementares ao texto - uma tabela com as questões e dois gráficos - estão no próprio texto. A base utilizada pela pesquisa são os microdados do Enem, disponibilizados pelo Inep. Foi descrito o passo-a-passo do processo, com as funções utilizadas nas etapas de seleção de itens suspeitos ou inadequados (*psych::biserial, mirt::itemfit, mokken::check.monotonicity* e *mirt::coef*). Além disso, foram usados hiperlinks para complementar o conteúdo. Em sua estrutura, o texto começa explicando o método utilizado de maneira resumida, detalha que tipos de dados foram usados (e também aqueles que foram descartados da análise, explicando os motivos), posteriormente, há uma tabela com todas as questões analisadas, com links para acessá-las, detalhamento do cálculo de parâmetros e da seleção de itens suspeitos de serem inadequados e, por fim, o processo de classificação de itens como inadequados, com demonstração a partir de dois gráficos de funções.

Analise_enem_2[20], sobre “90% da nota do Enem é influenciada por fatores econômicos e culturais, indica análise feita por GaúchaZH com uso de inteligência artificial”: o texto menciona que o processo escolhido foi a árvore de decisão, sem explicações da razão de escolha desse processo. O texto explica que os arquivos no repositório contém *scripts* de carga, tratamento e modelagem. O arquivo com o conjunto de treinamento é disponibilizado em uma pasta no Google Drive, mas sem possibilidade de acesso, na tentativa da autora. Não é identificada a base de dados usada como fonte e não há descrição do passo-a-passo do processo. A ferramenta mencionada é o Jupyter Notebook e não há hiperlinks.

Classificação de Logradouros Brasileiros por Gênero[21], sobre “Rua: substantivo (ainda) masculino”: nesse caso, o texto diz que foi feita uma classificação de logradouros brasileiros. Não há descrições das razões para as escolhas feitas ou descrição dos arquivos presentes no repositório. Descreve-se que foram usados os bancos de dados IBGE Nomes e dados dos logradouros brasileiros. O passo-a-passo é descrito como forma de instruções para reproduzir o processo desenvolvido pela equipe. Há menção às ferramentas utilizadas e um

[20] Disponível em <https://github.com/leonardojs1981/analise_enem_2>. Acesso em: 03/04/2022.
[21] Disponível em <https://github.com/generonumero/logradouros>. Acesso em: 03/04/2022.

hiperlink para uma página sobre por que os dados dos logradouros brasileiros não são abertos, mas deveriam ser. A estrutura inicia por uma descrição resumida do método, uma explicação de que os dados sobre logradouros foram comprados, seguida de instruções para reproduzir os passos desenvolvidos pela equipe.

Digging into the Mining Arc[22], sobre "Explorando o Arco Mineiro": não nomeia especificamente os procedimentos utilizados nem as razões para as escolhas. Também não há descrição do conteúdo das pastas ou descrição de bases de dados utilizada como fonte. Há instruções para que o processo feito possa ser reproduzido, menção às ferramentas utilizadas e utilização de hiperlinks. Em sua estrutura, o texto inicia com a identificação do autor da reportagem, depois breve descrição da reportagem, então disponibiliza instruções de instalação e de como executar o programa no Docker e, por fim, algumas observações sobre o processo.

01. Salário Mínimo[23], sobre "A evolução do número de aposentados que recebem 1 salário mínimo": não especifica os processos desenvolvidos, mas descreve o conteúdo dos arquivos no repositório. No fim do texto há uma relação das fontes dos dados. As descrições de processo fornecidas são apenas para baixar o repositório. Não há menção a ferramentas utilizadas e há hiperlinks para a matéria, para a página que possibilita o download do material, para a coluna publicada que baseou as análises e para mais informações sobre o Nexo Políticas Públicas. Na estrutura, o texto inicia com uma descrição do repositório, das pastas e arquivos presentes, então oferece instruções para baixar os arquivos, apresenta o autor e as fontes dos dados e, por fim, faz uma breve explicação da parceria para a publicação dos gráficos e convida a saber mais.

O que 15 mil tweets revelam sobre seu candidato[24]: o texto conta que os dados para a matéria foram raspados a partir de uma API do twitter. Não foi descrito o porquê da escolha dos processos e o passo-a-passo foi descrito junto ao detalhamento do conteúdo das pastas e arquivos no repositório. O texto também menciona as ferramentas utilizadas, mas o hiperlink utilizado é apenas para a reportagem originada a partir da análise feita. Na estrutura, há primeiramente uma breve descrição do repositório, depois uma descrição dos arquivos na pasta *code* e

---

[22] Disponível em <https://github.com/InfoAmazonia/arco-mineiro>. Acesso em: 03/04/2022.
[23] Disponível em <https://github.com/Nexo-Dados/PoliticasPublicas/tree/main/01.SalarioMinimo>. Acesso em: 03/04/2022.
[24] Disponível em <https://github.com/estadao/o-que-15-mil-tweets-revelam-sobre-seu-candidato/tree/e589f168130334677af8a4ff41df100a3970b539>. Acesso em: 03/04/2022.

com que finalidade eles foram usados, seguida da descrição do diretório *data* e por fim do diretório *viz*.

A evolução do novo coronavírus no Brasil: linha do tempo dos municípios atingidos pela epidemia[25]: não foram nomeados os procedimentos desenvolvidos, não foi descrito o conteúdo de cada arquivo específico, mas apenas o conteúdo do repositório em geral. Foi indicada a utilização da base de dados Brasil.io, não foi descrito o passo-a-passo. A autora relatou o uso das bibliotecas Pandas, Matplotlib e GeoPandas e há hiperlinks para a reportagem, a base de dados e as bibliotecas utilizadas.

An-site[26], sobre o Atlas da Notícia: não há texto de descrição do repositório.

5.4.2 Repositórios no GitHub

Foram analisados também os arquivos com códigos-fonte compartilhados no GitHub com o objetivo de entender os processos desempenhados para chegar-se ao resultado final em cada um dos casos. Para isso, buscou-se todos os arquivos do repositório que contêm *scripts* para uma análise mais detalhada do que foi feito, de que linguagens e ferramentas foram utilizadas e, também, se há comentários e indicações que ajudam a compreender o que foi feito.

A análise considerou as classificações dos processos de Big Data referenciadas por Gulia e Ratra (2019), **mineração** e **análise**, sendo as etapas da **análise** a *coleta*, *armazenamento*, *extração e filtragem* e *limpeza e validação*. Além de etapas de apresentação ou *visualização* dos dados. Os trabalhos não percorreram todas essas etapas, assim, serão analisadas apenas as que se aplicam a cada um deles. Além disso, a análise é feita a partir do material divulgado, o que pode excluir processos executados, mas não divulgados no GitHub, e, em alguns casos, desconsidera etapas de menor relevância para o resultado final.

Nomes mais registrados no Brasil em 2021: Publicado na conta do jornalista de dados Lucas Thaynan, um dos autores da reportagem, o repositório contém seis arquivos. No primeiro deles está a GNU General Public License v3.0, um tipo de licença que garante que qualquer um pode usar o software da forma que

---

[25] Disponível em <https://github.com/juditecypreste/linha-do-tempo-cidades-infectadas-com-coronavirus-no-Brasil>. Acesso em: 03/04/2022.
[26] Disponível em <https://github.com/voltdatalab/atlas-noticia>. Acesso em: 03/04/2022.

preferir, mas mantém os direitos do autor de forma que outros não podem apropriar-se daquele código-fonte e torná-lo privado. Depois há um arquivo com o texto que descreve o repositório, seguido por três arquivos em formato .csv, dois com produtos da raspagem, com ranking por estado e por município, e um com as coordenadas dos municípios. Por fim, há um arquivo com o código-fonte utilizado.

O objetivo desse trabalho foi extrair do Portal Oficial do Registro Civil os nomes mais registrados em cada estado e município brasileiro. Para isso, a equipe fez uma **análise**, com as etapas de *armazenamento* e *extração e filtragem*, utilizando a linguagem Python. O Jupyter Notebook é utilizado para interpretar e executar o *script*[27], que é manejado com o uso do conjunto de ferramentas de análise de dados Pandas. Com esses recursos, foram extraídos do Portal Oficial do Registro Civil primeiramente os dados totais de nascimentos por estado, a partir desses dados foram encontrados os 50 principais nomes em cada um dos estados e, enfim, os dados por município.

No arquivo com o código, todas as etapas estão claramente diferenciadas por meio de subtítulos descritivos, o que facilita a compreensão dos processos desempenhados.

Aplicativo sobre pensionistas do Fiquem Sabendo: este repositório foi publicado em uma conta própria da Fiquem Sabendo. Ele contém seis arquivos principais. Entre eles, há o texto de descrição do repositório, os dados resumidos do Portal de Transparência em formato .csv, instruções para executar a interface gráfica do aplicativo, microdados[28] com informações pessoais dos pensionistas, uma reprodução da consulta para a extração dos dados resumidos do banco de dados, seguida por um arquivo com instruções para visualizações intermediárias utilizadas na consulta e, por fim, no último arquivo está a consulta para a extração dos dados individualizados do último mês.

O objetivo desses processos foi tornar mais acessíveis ao público os dados divulgados sobre pensionistas, colocando-os numa interface interativa em que o usuário pode filtrar e pesquisar por informações específicas. Para isso, ele faz **análise** dos dados, primeiramente pré-processando, para a *limpeza* por meio de código-fonte em Python, os arquivos encontrados por mês no Portal da Transparência. Esse pré-processamento tem instruções disponíveis em um segundo

[27] Conjunto de instruções no software para execução de uma tarefa.
[28] Menor fração de um dado coletado em uma pesquisa.

repositório, postado na conta do GitHub de um dos colaboradores. O resultado desse processo é uma versão resumida dos dados com todas as combinações de tipos de beneficiário, órgão e cargo do servidor instituidor, quantidade de vínculos mantidos pelos pensionistas e estatísticas de valores pagos a pensionistas de cada uma das combinações. Esses arquivos, em formato .csv, são *armazenados* no sistema de gerenciamento de dados PostgreSQL, ferramenta com capacidade de armazenar dados de diferentes meses onde, a partir de consultas utilizando a linguagem SQL, os dados são *filtrados* para originar dois arquivos .csv que são, efetivamente, aqueles usados no aplicativo, que apresenta uma *visualização* interativa desenvolvida em RMarkdown (formato de arquivo que facilita a criação de documentos dinâmicos em R), convertida em html e entregue ao público com ajuda do portal de hospedagem shinyapps.io. Essa interface, produto final dos processos, permite ao usuário filtrar os dados de diferentes maneiras e fazer *downloads*.

Os arquivos têm subtítulos que descrevem cada etapa e nomeiam os resultados das extrações aplicadas à base de dados inicial.

Quadro 3 - Etapas de produção do aplicativo

Etapa	Ferramenta	Resultado
Download mês a mês	Manualmente	Arquivo zip com 3 planilhas
Limpeza e pré-processamento	Script Python	Arquivos .csv zipados
Inclusão do .csv no PostgreSQL	Comando no PostgreSQL	Banco de dados do Postgre populado
Extração do Postgre	Arquivos .sql disponíveis no repositório	2 arquivos .csv (microdados e consolidados)
Envio dos dois .csv e do .Rmarkdown para o servidor do aplicativo	Manualmente	Aplicativo no ar

Fonte: a autora (2022)

Material e métodos da reportagem "Questões do Enem na mira de Bolsonaro são eficientes em testar conhecimento": essa postagem, da conta no GitHub do DeltaFolha, não conta com um repositório, mas apenas com a descrição do que foi feito, portanto, não é possível fazer uma análise mais detalhada dos processos. O arquivo já faz parte de um repositório que contém as metodologias de

três reportagens desenvolvidas pela Folha de São Paulo a partir de microdados do Enem.

O objetivo dessa publicação foi avaliar se questões alvo de polêmica no governo Bolsonaro foram eficientes ou não para testar o conhecimento dos alunos. Para isso, foi utilizado um modelo estatístico que "estima a chance de um candidato acertar uma questão dada a sua proficiência na prova" (DELTAFOLHA, 2021) a partir de três parâmetros que fazem parte da metodologia utilizada pelo Inep para corrigir e dar notas, a Teoria da Resposta ao Item (TRI). Esses parâmetros são:

> "Discriminação (mede se a questão consegue diferenciar os candidatos de acordo com o nível de conhecimento naquele tema), parâmetro de dificuldade (indica o nível de dificuldade daquela questão) e parâmetro de acerto casual (estima a chance do candidato acertar porque chutou)" (DELTAFOLHA, 2021).

No processo, para obter grupos com representações mais homogêneas de candidatos por habilidade, a equipe agrupou os alunos pela quantidade de acertos removendo grupos com menos de 50 alunos, incluindo todos os candidatos com grupos de menos de 500 alunos e sorteando 500 candidatos em grupos com mais de 500 alunos. Para observar o comportamento da função, os parâmetros foram adicionados à função mirt (Multidimensional Item Response Theory), do pacote mirt da linguagem R, com suas respectivas distribuições, médias logarítmicas e desvios padrões, alpha e beta. Os itens suspeitos de serem inadequados foram identificados com as funções *psych::biserial* "para identificar itens que tenham seu padrão de respostas pouco correlacionado com o resto da prova, o que indicaria que o item em questão não tem sua resposta correta claramente determinada pela proficiência" (DELTAFOLHA, 2021); *mirt::itemfit* "para verificar se o modelo utilizado está prevendo de forma satisfatória a probabilidade de acertos no item segundo proficiência do candidato" (DELTAFOLHA, 2021); *mokken::check.monotonicity* "para identificar se em algum ponto da escala a probabilidade de acerto no item decresce quando a proficiência do candidato aumenta" (DELTAFOLHA, 2021); e *mirt::coef* "para detectar itens com pouca capacidade de distinguir alunos segundo sua proficiência" (DELTAFOLHA, 2021).

Ao final da descrição, há exemplos de funções resultantes do modelo, uma considerada suspeita de inadequada e outra claramente inadequada, porque não

obedece ao modelo esperado (que representa o maior acerto de alunos mais bem preparados).

Análise para "90% da nota do Enem é influenciada por fatores econômicos e culturais, indica análise feita por GaúchaZH com uso de inteligência artificial": este repositório não está em uma conta da GaúchaZH ou de algum dos jornalistas que participou da produção da matéria, mas sim do cientista de dados Leonardo Sales, que desenvolveu um modelo para prever a nota do Enem, no qual a equipe se baseou para produzir o modelo que originou essa matéria. O cientista também foi consultado pelos repórteres durante o processo de adaptação do modelo, utilizado em nível nacional. O repositório contém seis arquivos com os *scripts* das etapas desenvolvidas no processo.

O objetivo do trabalho foi investigar o quanto fatores socioeconômicos influenciam na nota do Enem. Com esse foco, o cientista de dados utilizou a linguagem Python na plataforma Jupyter Notebook primeiramente para *armazenar* um arquivo .csv com dados disponibilizados pelo Inep, montando um dataframe[29] Pandas com esses dados. O passo seguinte, depois de uma *filtragem* dos dados, foi salvar o dataframe com pickle[30], uma biblioteca para objetos python. Depois foi empreendida uma *limpeza* para retirar os alunos sem nota em todas as cinco áreas, baixando o arquivo de 1.583.756 para 1.346.866 linhas. Posteriormente são criadas colunas para médias, uma para a média nas provas objetivas e outra com a média geral (que inclui a redação), para estimar a renda familiar e per-capita do aluno a partir do piso da faixa de renda e duas contendo os valores 0 ou 1 (o que representa falso ou verdadeiro), discriminando os alunos que tiraram acima de 600 e os que tiraram acima de 700.

O passo seguinte foi sobrepor o arquivo pickle salvo anteriormente, assim, adicionando as novas colunas. Então são encontradas 30.662 escolas com a agregação[31] por escola com o cálculo das colunas:

```
"'media_objetivas_aluno_media',          'media_objetivas_aluno_mediana',
'media_geral_aluno_media',                  'media_geral_aluno_mediana',
'nota_matematica_media',                     'nota_matematica_mediana',
'nota_redacao_media',                           'nota_redacao_mediana',
'renda_per_capita_media',                    'renda_per_capita_mediana',
```

[29] Estrutura que organiza dados em uma tabela de colunas e linhas.
[30] Módulo utilizado para serializar e desserializar estruturas em Python.
[31] Tipo de associação em que o objetivo é demonstrar que informações de um objeto precisam ser complementadas por outro objeto.

'numero_de_alunos_enem', 'num_alunos_acima_600', 'num_alunos_acima_700'" (LEONARDOJS1981, 2018).

O passo seguinte foi executado a partir de dados do censo escolar 2016. Para cada região do Brasil ele coloca os dados do .csv em *dataframe*; *filtra* os dados, deixando apenas as colunas de código da entidade, tipo de etapa do ensino e matrícula do aluno; faz uma *limpeza* de duplicações de matrícula na mesma escola; e agrega o código da escola com a etapa de ensino para contar a quantidade de matrículas por combinação. O resultado desses processos é então colocado em um *dataframe* no qual são *filtradas* as matrículas totais de ensino médio por escola. O resultado é um *dataframe* com as colunas código da escola, quantidade de matrículas no ensino médio e quantidade total de matrículas.

No *script* seguinte, quarto arquivo do repositório, os dados tratados são do censo de docentes 2017. Ele *filtra* esses dados, identificando as escolas que tiveram alunos no ENEM 2017, utilizando, para essa identificação, os dados agregados anteriormente. A partir do processamento desses dados, chega-se a um *dataframe* com 5.780.908 docentes, que têm suas colunas de escolaridade convertidas em valores numéricos e ao qual são adicionadas as colunas de professores que têm curso de magistério, que lecionam no ensino médio e de qualificação dos professores (atribuindo pesos a cada tipo de qualificação). O arquivo que resulta desses processos é então *filtrado* para exibir apenas os professores que dão aula no ensino médio e salvo em *pickle*. Em seguida, os nomes duplicados de professores (que dão aula em mais de uma turma) são eliminados, o que reduz os dados para 605.941 docentes. Por fim, os dados são agregados por escola, o que resulta em 26.743 linhas.

No *script* seguinte são incluídas informações do censo escolar, cruzadas com os dados agregados de professores por escola, matrículas por escola e dados do Enem. Para eliminar pontos fora da curva, o *script* faz uma *filtragem*, eliminando escolas privadas com entre uma e três salas, menos de 30 alunos no total e que ficaram entre as 10% melhores (o que seriam “ilhas de excelência”). Em seguida são montadas *visualizações* em gráficos para relacionar a média geral dos alunos com as características da escola e em um gráfico de distribuição de média geral por aluno com microdados.

Depois são criados indicadores de estrutura das escolas, que são incluídos no *dataframe* montado, junto com os dados das escolas e microdados dos alunos.

Assim, para cada aluno da tabela de microdados do Enem são incluídas as informações selecionadas nos arquivos de *script* 3 (censo de alunos), 4 (censo de docentes) e 5 (censo de escolas). O resultado é um *dataframe* com 89 colunas, salvo em formato *pickle*.

Depois de executar algumas transformações no *dataframe* para facilitar a análise, o cientista de dados cria o indicador "estudou em escola pública estadual ou municipal". Essas operações marcam o final da etapa de pré-processamento de dados e o início do processo de **mineração**, com a montagem de uma *árvore de decisão*, ação que envolve *aprendizado de máquina*. Para isso, os atributos categóricos (em texto) são convertidos em atributos binários. Então, é usado o modelo de árvore de decisão DecisionTreeRegressor.

Nessa etapa, primeiramente ele identifica entre os 134 atributos menos correlacionados os 30 mais correlacionados com a média geral (assumindo que esses atributos vão influenciar mais na média geral do aluno), define parâmetros para limitar o tamanho da árvore de decisão e divide o conjunto de dados em dois, 90% para treinamento e 10% para teste do aprendizado de máquina. Depois é feita uma previsão com os dados de teste, que é comparada com os dados reais, observando-se que o erro médio absoluto da previsão é de 59,86 pontos.

Por fim, o último estágio do processo é uma função Python capaz de prever a nota de um aluno no Enem a partir dos 30 atributos mais correlacionados.

Nesse repositório, o processo é parcialmente descrito em comentários no código-fonte e os *outputs* ao longo do processo podem ser verificados.

Classificação de Logradouros Brasileiros por Gênero: publicado na conta da Gênero e Número, o repositório tem duas pastas, uma com todos os primeiros nomes de logradouros e outra com as diferentes possibilidades de grafia com cada nome, e arquivos com a GNU General Public License v3.0, o código-fonte e instruções para instalar as dependências do Python necessárias.

Os processos foram executados para classificar os logradouros brasileiros de acordo com o gênero. Utilizando a linguagem Python para o processamento de dados de registro civil do IBGE, a equipe verificou a frequência com que cada um dos nomes dos logradouros (e suas variações) apareciam nos registros como femininos ou masculinos.

O código é curto, não há comentários de instruções nele.

Digging into the Mining Arc: publicado na conta do próprio InfoAmazonia, este repositório contém o código-fonte de toda a interface projetada especialmente para essa reportagem. O repositório contém 13 arquivos com *script* e um com a licença GNU General Public License v3.0.

A interface, feita com JavaScript (98,2%) e HTML, inclui menu interativo, vídeos e fotos, gráficos e imagens panorâmicas. As instruções dos códigos-fonte foram projetadas para pessoas já familiares à plataforma Docker.

Gráficos em "A evolução do número de aposentados que recebem 1 salário mínimo": o repositório foi publicado na conta Nexo-Dados e contém o arquivo com o código utilizado para gerar os gráficos, a tabela de onde foram retirados os dados, uma pasta com os gráficos em formato pdf e outra com os gráficos em formato png.

O objetivo do trabalho foi criar visualizações de dados a respeito dos aposentados que recebem um salário mínimo. O resultado são quatro visualizações: um gráfico de barras com a evolução do número de pessoas que recebem aposentadoria; um segundo gráfico de barras com a evolução da porcentagem de aposentados que recebem um salário mínimo; um gráfico de área com o valor do salário mínimo ao longo do tempo; e um último gráfico de barras com a valorização real do salário mínimo. A linguagem utilizada para desenvolver as *visualizações* foi o R, com o uso das bibliotecas *tidyverse*, *readxl* e *lubridate* e o pacote com funções próprias do Nexo *nexo.utils*.

O arquivo com os códigos-fonte inclui comentários para diferenciar os processos definidos para cada uma das visualizações.

O que 15 mil tweets revelam sobre seu candidato: o diretório, publicado na conta do Estadão no GitHub, contém, além da licença GNU General Public License v3.0, três pastas, uma com os arquivos dos códigos para capturar tweets e para a análise geral e os resultados desses processos; outra com todos os tweets capturados, os tweets capturados no recorte temporal da análise, valores de frequência e unicidade para as palavras de cada candidato e um arquivo com todos os candidatos em apenas uma planilha; e uma última com a função que projeta os gráficos da matéria.

O objetivo da **análise** foi identificar quais foram as palavras mais características usadas nos tweets de cada candidato às eleições presidenciais de 2018. Primeiramente, foi utilizada uma API do twitter e o módulo Tweepy do Python,

no Jupyter Notebook e utilizando o Pandas, para *coletar* os tweets mais recentes dos candidatos. Então, foram definidas credenciais de acesso e os tweets foram transferidos para arquivos de texto. O passo seguinte foi fazer uma *limpeza* dos dados criando um *dataframe* para adicionar os arquivos de cada candidato, excluindo os retweets e os tweets do período anterior ao selecionado para a análise e duplicatas. Depois de outros processos de tratamento, foi criado um *dataframe* para cada candidato, o *dataframe* com os tweets de todos os candidatos a partir da data selecionada foi salvo em .csv para ser usado posteriormente na visualização.

Em seguida foi realizada mais uma *limpeza*, removendo hashtags, menções e links, preparando os dados para o processo de tokenização, que separa cada elemento linguístico e envolve a criação de um *dataframe* com percentuais de uso de cada palavra.

Depois de uma *filtragem* para excluir palavras sem significados que interessam à análise (como preposições e artigos), com menos de três letras e usadas comumente por todos os candidatos, operações foram executadas para descobrir quantas vezes cada palavra foi utilizada e foi feito um cálculo da proporção de uso de cada palavra em relação às demais, por candidato, e comparação dessas proporções entre candidatos. Por fim, os dados foram salvos em um dicionário.

Completada a análise, os resultados de cada candidato foram concatenados em um arquivo único.

O repositório também contém os códigos-fonte utilizados para gerar as visualizações dos dados, estes em JavaScript.

Nos arquivos deste repositório, um comentário acompanha cada passo tomado, além dos títulos que agrupam os scripts por etapa do processo.

A evolução do novo coronavírus no Brasil: linha do tempo dos municípios atingidos pela epidemia: este repositório está na conta do GitHub da jornalista Judite Cypreste, autora da matéria. Ele contém um arquivo com o código-fonte utilizado e outro com as coordenadas das cidades brasileiras.

A finalidade do trabalho foi analisar a evolução do coronavírus no Brasil entre 26 de fevereiro e 1 de maio de 2020 numa visualização em mapa. Para isso foi utilizada a linguagem Python no Jupyter Notebook no cruzamento dos dados de infecções por coronavírus com os dados de latitude e longitude dos municípios, gerando um arquivo .csv para cada dia analisado. O resultado disso foi então utilizado para construir o mapa visto na matéria.

O arquivo com o código-fonte, como está misturado aos dados de infecções, é grande demais para ser aberto no navegador, então precisa ser baixado para visualização em html ou no Jupyter Notebook. A repórter também descreve as etapas com comentários no *script*.

Material sobre o Atlas da Notícia: armazenado no repositório do Volt Data Lab, o material contém uma série de scripts em HTML do site do Atlas da Notícia, além de imagens, gráficos e outros tipos de arquivo.

Na tabela com os dados dos veículos brasileiros, fruto de pesquisa ou informação voluntária do próprio veículo, é possível filtrar as informações de diferentes formas, copiar a tabela e baixar os dados.

5.5 Análise por categorias

Na etapa seguinte prevista pela metodologia adotada nesta pesquisa, foram elaboradas categorias segundo as quais o *corpus* é analisado, tanto em relação aos *scripts* disponibilizados quanto referindo-se ao material publicado como um todo. Com o objetivo de responder o problema de pesquisa e analisar de forma mais ampla, as categorias criadas para este trabalho foram *transparência*, *profundidade*, *abertura*, *relevância* e *interatividade*.

5.5.1 Transparência

Diz respeito à divulgação dos procedimentos e métodos seguidos para construir uma peça jornalística. Nesse quesito, a principal pergunta é se há meios e com que facilidade é possível que o leitor reproduza os processos desempenhados para a reportagem. Nesse aspecto, baseando-se em critérios de análise de Gehrke (2018b), é verificado se há algum tipo de descrição dos procedimentos e/ou link para a página no GitHub na publicação, se as instruções para repetir os processos são claras, se as fontes dos dados são divulgadas e se há correções e atualizações no material.

Quadro 4 - Descrição dos procedimentos utilizados em cada uma das peças

Peça	Divulgação dos procedimentos
Conheça os nomes mais populares de 2021 em cada estado brasileiro	Ao longo do texto é dito que foi feita uma análise e ao final da matéria é disponibilizado um link para o perfil no GitHub onde os dados foram disponibilizados.
Fiquem Sabendo lança aplicativo para consulta de remuneração de pensionistas	Nesse material, que anuncia e descreve o aplicativo, há menções breves ao processo realizado e não há link para o GitHub.
Questões do Enem na mira de Bolsonaro são eficientes em testar conhecimento	Menciona e explica constantemente ao longo da matéria os métodos utilizados, inclusive utilizando visualizações para isso. Ao final há um resumo da metodologia com link para mais informações no GitHub.
90% da nota do Enem é influenciada por fatores econômicos e culturais	Já no início explica brevemente os métodos utilizados. Ao final há um menu tipo sanfona com informações sobre o que motivou a análise, como a matéria foi apurada e quais foram os cuidados tomados para que a análise fosse justa, junto ao link para o código aberto do modelo no qual a equipe se baseou, na conta de Leonardo Sales no GitHub
Rua: substantivo (ainda) masculino	Há apenas uma referência a programação e análise de dados, nos créditos do vídeo, sem explicações sobre metodologia
Explorando o Arco Mineiro	Nesse caso, os dados não são o que motivou a história. Eles são usados para complementar o texto e não há explicações sobre como foram tratados e transformados nos gráficos vistos na reportagem
A evolução do número de aposentados que recebem 1 salário mínimo	Não fala sobre a metodologia, mas disponibiliza o link do repositório do GitHub no final referindo-se a ele como meio de acesso aos dados utilizados
O que 15 mil tweets revelam sobre seu candidato	No início do texto a metodologia já começa a ser explicada e, no fim, há uma explicação mais detalhada, mas sem link para o GitHub
Coronavírus avança para o interior do Brasil; veja evolução em mapa	O texto diz que o UOL analisou dados e preparou uma linha do tempo da disseminação do vírus e fala das fontes dos dados
Atlas da notícia	Não mostra como os dados foram obtidos e processados

Fonte: a autora (2022)

A partir do quadro, observa-se que duas reportagens (em verde), as que analisam dados do Enem, optaram por descrever os métodos ao longo do texto, detalhar ao final e complementar com um link para o GitHub. Essa escolha pode ter ocorrido porque as análises desempenhadas nesses casos, as duas estatísticas, uma delas sabidamente feita com auxílio de machine learning, são complexas, demandantes e mais dificilmente entendidas pelo público, além de abordarem temas

relativamente polêmicos. Observa-se também que os dois veículos que optaram por essa forma mais completa de divulgação da metodologia, GaúchaZH e Folha de São Paulo, fazem parte de grandes conglomerados de mídia.

Outras duas peças (em amarelo) optaram por falar dos processos de duas formas diferentes. Na peça de O Estado de São Paulo, a metodologia é descrita resumidamente no início e mais detalhadamente no final, mas sem links para o repositório do GitHub. Esta também é uma análise com maior nível de complexidade e elaborada por um veículo que faz parte de um grande conglomerado de mídia. Já a reportagem sobre nomes mais comuns, da Agência independente Tatu, opta por mencionar brevemente os processos no texto e depois disponibilizar o link para o GitHub, mas não sem antes enfatizar no próprio texto que é um veículo comprometido com a transparência.

Uma outra reportagem (em azul), do Nexo Jornal, não faz referência aos processos no texto mas aponta para o repositório do GitHub, que contém mais detalhes, e outras três (em roxo) apenas mencionam a análise e processamento dos dados brevemente. As demais não mencionam esse processo.

Quadro 5 - Clareza na descrição no GitHub

Peça	Descrição no GitHub
Conheça os nomes mais populares de 2021 em cada estado brasileiro	Texto indica a ferramenta utilizada e descreve o conteúdo gerado como resultado (e em que arquivos do repositório ele se encontra), o banco de dados utilizado e o motivo da utilização do método escolhido. No script há títulos para cada processo realizado e comentários descrevendo as etapas. Também é divulgada a fonte dos dados com link para acesso.
Fiquem Sabendo lança aplicativo para consulta de remuneração de pensionistas	O texto menciona os métodos utilizados e trabalho desenvolvido para cada funcionalidade, detalha etapas de processamento dos dados, descreve o conteúdo dos arquivos encontrados no repositório, discorre sobre o processo para conseguir que os dados fossem disponibilizados e destaca pontos aos quais o leitor deve se atentar ao acessar o repositório. No script há alguns comentários relativamente vagos sobre as etapas. A fonte dos dados, com hiperlink, também é disponibilizada.
Questões do Enem na mira de Bolsonaro são eficientes em testar conhecimento	Especifica métodos e ferramentas utilizadas, apresenta os dados com auxílio de uma tabela e expõe os resultados com ajuda de dois gráficos. Não há script a ser comentado. A fonte dos dados é disponibilizada, com hiperlink para ela.
90% da nota do Enem é influenciada por fatores	Especifica o processo utilizado, descreve o conteúdo nos arquivos do repositório. O arquivo com o conjunto de treinamento é disponibilizado em uma pasta no Google Drive, mas sem possibilidade de acesso, na tentativa da autora. Não é identificada a base de dados usada como fonte e não há

econômicos e culturais	descrição do passo-a-passo do processo. Nos scripts há títulos, subtítulos e comentários identificando etapas. Também não há hiperlinks.
Rua: substantivo (ainda) masculino	Identifica processos, ferramentas e bancos de dados utilizados, com link para o banco que é público. Não há descrições das razões para as escolhas feitas ou descrição dos arquivos presentes no repositório. A estrutura inicia por uma descrição resumida do método, uma explicação de que os dados sobre logradouros foram comprados, seguida de instruções para reproduzir os passos desenvolvidos pela equipe. No script não há comentários
Explorando o Arco Mineiro	Não há descrição do conteúdo das pastas ou descrição da base de dados utilizada como fonte. Há instruções para que o processo feito possa ser reproduzido, menção às ferramentas utilizadas e utilização de hiperlinks. Há raros comentários nos scripts
A evolução do número de aposentados que recebem 1 salário mínimo	Não especifica os processos desenvolvidos, mas descreve o conteúdo dos arquivos no repositório. No fim do texto há uma relação das fontes dos dados. As descrições de processo fornecidas são apenas para baixar o repositório. Não há menção a ferramentas utilizadas e há hiperlinks para a matéria, para a página que possibilita o download do material, para a coluna publicada que baseou as análises e para mais informações sobre o Nexo Políticas Públicas. As etapas do processo são demarcadas com comentários no script.
O que 15 mil tweets revelam sobre seu candidato	O passo-a-passo foi descrito junto ao detalhamento do conteúdo das pastas e arquivos no repositório. O texto também menciona as ferramentas utilizadas, mas o hiperlink utilizado é apenas para a reportagem originada a partir da análise feita. Cada etapa script é detalhadamente descrita.
Coronavírus avança para o interior do Brasil; veja evolução em mapa	Não foram nomeados os procedimentos desenvolvidos, não foi descrito o conteúdo de cada arquivo específico, mas apenas o conteúdo do repositório em geral. Foi indicada a utilização da base de dados Brasil.io, não foi descrito o passo-a-passo. A autora relatou o uso das bibliotecas Pandas, Matplotlib e GeoPandas e há hiperlinks para a reportagem, a base de dados e as bibliotecas utilizadas. O script é brevemente comentado.
Atlas da notícia	Não há descrição do conteúdo. Há poucos comentários no código

Fonte: a autora (2022)

Desses, o mais difícil de ser reproduzido é “Questões do Enem na mira de Bolsonaro são eficientes em testar conhecimento”, pois o jornal não compartilha o código-fonte. *Scripts* sem muitas camadas de pré-processamento, como em “Conheça os nomes mais populares de 2021 em cada estado brasileiro” e “Rua: substantivo (ainda) masculino” são mais fáceis de serem reproduzidos. O material que se destaca pelo detalhamento é “O que 15 mil tweets revelam sobre seu candidato”, que descreve cada passo em comentários no *script*. “Explorando o Arco Mineiro” e “Atlas da Notícia” optaram por compartilhar scripts de todo o site.

5.5.2 Profundidade

Esta categoria serve para avaliar os conteúdos de acordo com o nível de contextualização e explicação dos fenômenos analisados. A seguir analisa-se os dados e o complemento aos dados em cada uma das peças.

Quadro 6 - Capacidade de contextualização e aprofundamento das peças

Peça	**Como contextualiza e se aprofunda no assunto**
Conheça os nomes mais populares de 2021 em cada estado brasileiro	Descrição dos resultados da análise com atenção a padrões e singularidades e explicação do fenômeno observado por um antropólogo
Fiquem Sabendo lança aplicativo para consulta de remuneração de pensionistas	Revela processo que levou a divulgação dos dados
Questões do Enem na mira de Bolsonaro são eficientes em testar conhecimento	Com um histórico dos ataques feitos por políticos às questões, verificação a respeito, também, do desempenho de candidatos transexuais e especialistas falando a respeito do método de análise e questões políticas
90% da nota do Enem é influenciada por fatores econômicos e culturais	Mostra médias de notas obtidas em alguns estados, além de explicações e exemplos sobre a relação entre esforço, condições socioeconômicas e nota, com uma fonte especializada e depoimentos de uma aluna de Psicologia moradora do bairro Lomba do Pinheiro
Rua: substantivo (ainda) masculino	Mostrando as características da nomeação dos logradouros brasileiros ao longo da história e, mais especificamente, em alguns estados
Explorando o Arco Mineiro	Com depoimentos de mineiros, empresas, acadêmicos, indígenas, políticos e ativistas sobre a disputa por áreas de mineração na Venezuela. O conteúdo inclui texto, vídeos, fotos e gráficos numa interface criada especialmente para a reportagem
A evolução do número de aposentados que recebem 1 salário mínimo	Os gráficos não acompanham quase nenhum texto, o que há de contexto é oferecido a partir do link para outra matéria do Nexo, que motivou a análise
O que 15 mil tweets revelam sobre seu candidato	A reportagem faz comentários breves relatando os achados mais relevantes de cada candidato
Coronavírus avança para o interior do Brasil; veja evolução em mapa	Faz um histórico e descrição da situação brasileira em relação ao vírus até o momento de publicação
Atlas da notícia	A tabela em si contém os veículos e desertos de notícias com suas respectivas possibilidades de filtros, sem interpretações. Mas, na página do Atlas da Notícia são encontradas diversas matérias com aprofundamento e interpretações do conteúdo

Fonte: a autora (2022)

Primeiramente, observa-se que os conteúdos “Conheça os nomes mais populares de 2021 em cada estado brasileiro”, “Questões do Enem na mira de Bolsonaro são eficientes em testar conhecimento”, “90% da nota do Enem é influenciada por fatores econômicos e culturais” e “Explorando o Arco Mineiro” optaram por fazer entrevistas e incluir declarações de fontes para explicar os fenômenos observados. Fontes especializadas foram citadas em todos esses materiais e dois deles, “90% da nota do Enem é influenciada por fatores econômicos e culturais” e “Explorando o Arco Mineiro”, também utilizaram “pessoas comuns” como fontes.

Também há as peças que fazem pouca ou nenhuma interpretação, como o “Atlas da Notícia” e “A evolução do número de aposentados que recebem 1 salário mínimo”. Outros optam por descrições mais simples, que são “Fiquem Sabendo lança aplicativo para consulta de remuneração de pensionistas” e “O que 15 mil tweets revelam sobre seu candidato”.

“Questões do Enem na mira de Bolsonaro são eficientes em testar conhecimento” traz, também, uma análise de dados complementar ao conteúdo e um histórico dos ataques feitos ao Enem. “Rua: substantivo (ainda) masculino” também opta por trazer um contexto histórico para complementar a análise dos dados, assim como “Coronavírus avança para o interior do Brasil; veja evolução em mapa”.

5.5.3 Abertura

Essa categoria se refere ao comprometimento dos veículos jornalísticos analisados com uma política de dados abertos e ações de transparência. A manutenção de uma conta no GitHub do próprio veículo e a preocupação com obtenção de licenças que assegurem que todos possam fazer uso do código podem ser considerados indícios de que a organização está preocupada com essas questões.

Quadro 7 - Compromisso com a abertura do conteúdo

Peça	Utilização de licença para conteúdo no GitHub	Processo divulgado na conta do próprio veículo no GitHub
Conheça os nomes mais populares de 2021 em cada estado brasileiro	Sim, a GNU General Public License v3.0	Não, foi divulgado na conta do repórter Lucas Thaynan
Fiquem Sabendo lança aplicativo para consulta de remuneração de pensionistas	Não	Sim
Questões do Enem na mira de Bolsonaro são eficientes em testar conhecimento	Não	Sim, foi divulgado na conta do DeltaFolha, pertencente à Folha de São Paulo
90% da nota do Enem é influenciada por fatores econômicos e culturais	Não	Não. O processo está na conta do cientista de dados Leonardo Sales, que desenvolveu o método no qual a reportagem se baseou
Rua: substantivo (ainda) masculino	Sim, a GNU General Public License v3.0	Sim
Explorando o Arco Mineiro	Sim, a GNU General Public License v3.0	Sim
A evolução do número de aposentados que recebem 1 salário mínimo	Sim, a Creative Commons	Sim, na conta do Nexo Dados
O que 15 mil tweets revelam sobre seu candidato	Sim, a GNU General Public License v3.0	Sim
Coronavírus avança para o interior do Brasil; veja evolução em mapa	Não	Não, foi publicado na conta de Judite Cypreste, autora da reportagem
Atlas da notícia	Não	Sim

Fonte: a autora (2022)

A partir do quadro, é possível constatar que todos os materiais que utilizam licenças são de veículos independentes, exceto a reportagem de O Estado de São Paulo. Ao mesmo tempo, muitos veículos que têm contas próprias no GitHub não utilizam licença.

5.5.4 Relevância

Essa categoria avalia qual foi a importância da utilização de linguagens de programação para se chegar aos resultados observados nas peças jornalísticas. As duas reportagens que utilizaram modelos estatísticos para gerar conhecimento novo a partir dos dados, “Questões do Enem na mira de Bolsonaro são eficientes em testar conhecimento” e “90% da nota do Enem é influenciada por fatores econômicos e culturais”, estão entre as operações mais complexas. Também estão nesse grupo “O que 15 mil tweets revelam sobre seu candidato”, com várias etapas de pré-processamento e cálculo. Nessas reportagens, muito dificilmente seria possível chegar às conclusões e resultados sem a análise com auxílio de programação.

Enquanto isso, em “Conheça os nomes mais populares de 2021 em cada estado brasileiro”, “Rua: substantivo (ainda) masculino” e “Coronavírus avança para o interior do Brasil; veja evolução em mapa” não são executadas tantas operações e processos de análise, mas são analisadas informações sobre a população de todo o Brasil, o que também torna essencial o processamento por meio de *script*, sem o qual os assuntos e motivos principais de existência das peças não existiriam.

Por sua vez, em “Fiquem Sabendo lança aplicativo para consulta de remuneração de pensionistas” foram realizados diferentes processamentos dos dados para tornar o conteúdo acessível, assim, mesmo que o aplicativo não apresente análises dos dados, a programação foi fundamental. O aplicativo tem um objetivo similar ao “Atlas da Notícia”, este que aparentemente não teve necessidade de processar os dados de muitas formas diferentes para mostrar a informação, utilizando programação para desenvolver a interface que torna os dados mais acessíveis.

A reportagem “Explorando o Arco Mineiro” também não apresenta análises dos dados. Nela, a programação serve para enriquecer o conteúdo, possibilitando maior interatividade e facilidade na produção de gráficos. A programação também vem como uma forma de facilitar a produção e enriquecer o conteúdo em “A evolução do número de aposentados que recebem 1 salário mínimo”, em que ela é usada para a produção de gráficos.

5.5.5 Interatividade

Esse quesito diz respeito às visualizações de dados utilizadas nas peças jornalísticas. Essa interatividade pode ser usada para personalizar a visualização dos dados ou para tornar o conteúdo mais dinâmico e lúdico.

Quadro 8 - Interatividade

Peça	Elementos de interatividade
Conheça os nomes mais populares de 2021 em cada estado brasileiro	Mostra nomes dos respectivos estados e nomes mais registrados no território ao passar o cursor sobre o mapa
Fiquem Sabendo lança aplicativo para consulta de remuneração de pensionistas	Menu com diferentes filtros e operações que podem ser feitas com os dados, além de botões para download
Questões do Enem na mira de Bolsonaro são eficientes em testar conhecimento	As visualizações da matéria são estáticas, não há elementos interativos nelas
90% da nota do Enem é influenciada por fatores econômicos e culturais	Ao passar o cursor sobre os gráficos de barras as informações sobre cada um dos números aparecem resumidas em uma caixa
Rua: substantivo (ainda) masculino	A reportagem é em vídeo
Explorando o Arco Mineiro	Além do menu, a interface tem diversos recursos interativos, como imagens e vídeos que aparecem no lado direito da tela à medida que a página é rolada e uma foto panorâmica que pode ser manipulada para que o leitor enxergue de diferentes ângulos
A evolução do número de aposentados que recebem 1 salário mínimo	As visualizações não são interativas
O que 15 mil tweets revelam sobre seu candidato	No início da matéria as visualizações de alguns candidatos rolam horizontalmente, também é possível puxá-las com o próprio mouse. Além disso, para ver o texto completo de um candidato é necessário clicar num botão de expandir
Coronavírus avança para o interior do Brasil; veja evolução em mapa	É possível escolher entre “dar play” para ver a passagem dos dias no mapa automaticamente ou passar dia por dia manualmente
Atlas da notícia	É possível baixar dados, copiar tabelas e visualizar os dados com diferentes filtros

Fonte: a autora (2022)

De todas as peças analisadas, apenas três não têm nenhum tipo de interatividade na visualização e exibição dos dados. Nas peças “Conheça os nomes mais populares de 2021 em cada estado brasileiro”, “90% da nota do Enem é influenciada por fatores econômicos e culturais” as partes interativas não acrescentam informações, apenas repetindo o que já é mostrado sem necessidade de interação do leitor.

Enquanto isso, “Fiquem Sabendo lança aplicativo para consulta de remuneração de pensionistas” e o “Atlas da notícia” oferecem um conjunto mais amplo de informações. Próprias para servir de base para outros materiais jornalísticos, essas peças oferecem filtros e opções de download dos dados.

“Explorando o Arco Mineiro” é uma grande reportagem multimídia, com interação tanto em sua interface quanto no conteúdo propriamente dito, e “O que 15 mil tweets revelam sobre seu candidato” usa interação para tornar o conteúdo mais dinâmico e dar opções para o leitor de ocultar e mostrar informações. Por fim, “Coronavírus avança para o interior do Brasil; veja evolução em mapa” utiliza o recurso para que a pessoa possa investigar os dados na velocidade que preferir.

5.5.6 Visão geral

O uso de novas ferramentas tecnológicas no jornalismo oportuniza a produção de novos formatos de reportagem e descoberta de informações antes ocultas ou difícil de serem conseguidas. Com os processos e ferramentas como os encontrados nesta monografia (mostrados no quadro abaixo), é possível ir além de um jornalismo declaratório e expor o que, de fato, acontece, além das opiniões. Esse conjunto de fatores cria também uma cultura de cooperação entre jornalistas, que inclui noções de código aberto e da cultura hacker.

Quadro 9 - Principais processos, linguagens e ferramentas utilizados

Peça	Processos	Linguagens	Recursos
Conheça os nomes mais populares de 2021 em cada estado brasileiro	Análise; filtragem; extração	Python	Jupyter Notebook; Pandas
Fiquem Sabendo lança aplicativo para consulta de remuneração de	Extração; visualização; análise; limpeza	Python; SQL; R	PostgreSQL; shinyapps.io

pensionistas			
Questões do Enem na mira de Bolsonaro são eficientes em testar conhecimento	Análise estatística; limpeza	R	Pacote mirt; psych::biserial; mokken::check.monotonicity
90% da nota do Enem é influenciada por fatores econômicos e culturais	Limpeza; cruzamento; visualização; mineração (árvore de decisão/machine learning)	Python	Jupyter Notebook; Pandas; DecisionTreeRegressor
Rua: substantivo (ainda) masculino	Processamento	Python	--
Explorando o Arco Mineiro	Montagem de interface	JavaScript; HTML	Docker
A evolução do número de aposentados que recebem 1 salário mínimo	Visualização	R	Tidyverse; readxl; lubridate; nexo.utils
O que 15 mil tweets revelam sobre seu candidato	Coleta; visualização; análise; limpeza; tokenização; filtragem; concatenação	Python; JavaScript	API do Twitter; Tweepy; Jupyter Notebook; Pandas
Coronavírus avança para o interior do Brasil; veja evolução em mapa	Cruzamento; análise	Python	Jupyter Notebook; Pandas
Atlas da notícia	Material do site	HTML	--

Fonte: a autora (2022)

Pode-se observar que o processo de limpeza dos dados é o que mais se repete nos trabalhos, tendo sido utilizado em quatro deles. Cruzamento, filtragem, visualização e extração (sem contar o processo de “análise”, com definição mais genérica) também aparecem mais de uma vez. A linguagem mais utilizada é Python, que aparece seis vezes, seguida do R, presente em três trabalhos. A ferramenta mais utilizada é o Jupyter Notebook, o que reflete o fato de a principal linguagem ser Python. Os tipos de processo utilizados são variados, mas observa-se que na maioria das vezes o objetivo dos processos é facilitar a compreensão e outros processamentos de um conjunto relativamente grande de dados.

6 CONCLUSÃO

Num contexto em que há uma infinidade de informações disponíveis, organizar e tornar os dados compreensíveis nunca foi tão necessário. Observando-se essa necessidade, quanto mais ferramentas e noções de programação e estatística o jornalista tem, mais formas de descobrir informações e contar histórias estarão ao seu alcance. Nesse sentido, há iniciativas tanto de veículos independentes quanto de conglomerados de mídia de fazer um jornalismo especializado que utiliza essas ferramentas para ir além das declarações de fontes e até confirmar ou refutar percepções em debate na esfera pública.

O tema deste trabalho foi escolhido pensando nessas questões e a pesquisa foi motivada por um interesse em entender esses processos de apreensão da realidade por meio da utilização de ferramentas para desvendar informações escondidas em planilhas e sites da internet. Além disso, a pesquisa sobre esse tema ainda é limitada, em especial no Brasil, e é importante para a sociedade como um todo entender os métodos por meio dos quais informações são obtidas, o que dá relevância ao tema em questão.

Ao longo da pesquisa, foi observado que há uma vontade entre os autores das peças estudadas de documentar os processos e compartilhá-los, o que se encaixa na cultura *hacker* reproduzida no meio jornalístico, caracterizada por Träsel (2018). Em alguns momentos isso é feito com maior abertura e em outros alguns conteúdos são ocultados. Também observa-se diferentes níveis de preocupação em descrever o passo-a-passo, para deixar o material mais acessível e reprodutível. Observou-se também que a linguagem utilizada na maioria dos projetos é o Python e que os processos de análise, incluindo principalmente a filtragem, cruzamento e limpeza de dados, são os mais comuns. Também há outros métodos observados, como, por exemplo, processos de *machine learning*. Essas percepções contribuem para um maior entendimento dos métodos e para inspirar jornalistas a enxergar novas possibilidades na prática de jornalismo de dados.

Com relação ao problema de pesquisa, **“**como os jornalistas-programadores brasileiros exercem a transparência sobre os métodos que utilizam para obter informações a partir de bases de dados?”, foi constatado que na maioria dos trabalhos há esforços para compartilhar um material acessível no GitHub, em especial nas descrições dos métodos que acompanham os repositórios.

Também há considerável variedade nos processos. Enquanto alguns realizaram operações de cruzamento, limpeza e filtragem de dados para encontrar uma resposta mais rapidamente, sem precisar analisar a base de dados linha por linha, como no caso de “Conheça os nomes mais populares de 2021 em cada estado brasileiro” e “Rua: substantivo (ainda) masculino”, outros optam por análises estatísticas e processamentos para compreender padrões que seriam dificilmente identificados com precisão apenas com a observação humana, como em “Questões do Enem na mira de Bolsonaro são eficientes em testar conhecimento”, “90% da nota do Enem é influenciada por fatores econômicos e culturais” e “O que 15 mil tweets revelam sobre seu candidato”.

Também há ocasiões em que os dados são processados para torná-los mais acessíveis em formatos específicos de visualização e interação, o que acontece com ”Fiquem Sabendo lança aplicativo para consulta de remuneração de pensionistas” e “Coronavírus avança para o interior do Brasil; veja evolução em mapa”. Também há casos em que eles são coletados de uma base de dados e transformados em visualizações, como em “A evolução do número de aposentados que recebem 1 salário mínimo”.

Com relação às publicações no GitHub, na maioria das vezes há explicações detalhadas no texto que descreve o repositório, apesar de que, dentro do código-fonte, grande parte dos processos não costuma ser descrito por comentários. Assim, em seis dos dez trabalhos analisados, é possível compreender os processos utilizados antes de observar o *script*, apenas com a descrição textual feita no GitHub; e, em sete peças, a divulgação é feita na conta do veículo jornalístico, não pelo jornalista envolvido ou outra pessoa.

Em relação às linguagens de programação utilizadas, a que apareceu mais vezes foi Python, encontrada em seis dos dez trabalhos analisados. Além disso, também foram encontradas as linguagens R, HTML, JavaScript e SQL.

Também pode ser constatado que as abordagens acrescentam informações relevantes e que não poderiam ou seriam muito dificilmente encontradas de outras formas, principalmente nos casos em que há processos de aprendizado de máquina e análises estatísticas. Também foi observado que os processamentos são diversos, muitas vezes não tendo uma classificação específica e o processo que apareceu com mais frequência foi a limpeza de dados.

A partir desta monografia, é possível observar como o aprendizado a respeito de novas tecnologias e conhecimentos técnicos, tanto sobre linguagens de programação quanto matemáticos e estatísticos, fazem diferença na atividade jornalística. Esses conhecimentos possibilitam o acesso a uma infinidade de novas informações, já disponíveis na rede, que também necessitam de um olhar das ciências sociais para que se tornem acessíveis ao público e possam servir à sociedade no sentido de favorecer a democracia, os direitos humanos e chamar atenção para questões importantes.

As análises também indicam que a transparência no jornalismo pode ir muito além da revelação das fontes da informação. Exemplos disso são as peças “Atlas da Notícia” e “Explorando o Arco Mineiro”, que optaram por divulgar no GitHub os códigos-fonte de suas páginas na web quase na íntegra. Recursos como o GitHub, se melhor explorados, podem abrir possibilidades para uma colaboração nunca antes vista no jornalismo, tanto entre jornalistas, quanto dos jornalistas com o público. Essa colaboração pode envolver, por exemplo, projetos de investigação sobre questões globais, criados em conjunto por jornalistas de diferentes partes do mundo; *insights* trazidos por leitores; e conversas entre jornalistas em diferentes partes do Brasil e do mundo que estejam desenvolvendo projetos parecidos.

Quanto mais detalhada for a descrição dos métodos utilizados, maiores são as possibilidades de colaboração e maior o nível de transparência do material. Considerando que a reprodutibilidade dos processos é uma das características de um jornalismo transparente, ao criar um repositório no GitHub o autor pode pensar em como tornar o método o mais reprodutível possível para alguém com o mínimo de conhecimento técnico. Isso pode ser feito a partir de uma descrição detalhada de cada etapa do processo no texto que acompanha o repositório, seguida pela descrição do conteúdo de cada uma das pastas (de preferência nomeadas de forma facilmente identificável), identificação das linguagens e ferramentas utilizadas e, se necessário, instruções de como instalar os programas utilizados e como interpretar o conteúdo do repositório. Dentro do código-fonte, descrever o que cada comando faz e qual a sua função para atingir o resultado final por meio de comentários também confere mais facilidade de entendimento do conteúdo.

Essas observações são resultado de uma análise geral e superficial que pincelou diversos aspectos da questão. No futuro, cada processo encontrado

poderia ser melhor explorado, assim como os métodos de transparência das organizações jornalísticas que desenvolvem matérias de jornalismo de dados.

REFERÊNCIAS

AGÊNCIA TATU. **Sobre**, 2017. Página “Sobre”**.** Disponível em: <https://www.agenciatatu.com.br/sobre/>. Acesso em: 30/03/2022.

AISCH, Gregor. **Using Data Visualization to Find Insights in Data**. In: The Data Journalism Handbook 1, p. 139-148, 2011.

ANDERSON, C. W.; BELL, E.; SHIRKY, C. **Jornalismo pós-industrial: adaptação aos novos tempos**. Revista de jornalismo ESPM. São Paulo, Ano 2, N.5, Abr. Maio. Jun. 2013. p. 30-89.

ATLAS DA NOTÍCIA. **Sobre o Atlas da Notícia**, 2022. Página “Sobre o Atlas da Notícia”**.** Disponível em: <https://www.atlas.jor.br/institucional/sobre-o-atlas-da-noticia>. Acesso em: 30/03/2022.

BARBOSA, Suzana; ALBAN, Renato. **Convergência jornalística e uso de bases de dados no trabalho jornalístico**. Estudo do caso Correio. Universidade Federal da Bahia. Brasil, 2013.

BARBOSA, Suzana. **Jornalismo Digital em Base de Dados (JDBD). Um paradigma para produtos jornalísticos digitais dinâmicos**. (Tese de Doutorado). FACOM/UFBA, Salvador, 2007.

BATISTA, André Luiz França; FILHO, Aurélio Pajuaba Nehme; PIMENTEL, Daniel Ramos; MARTINS, Rodrigo Grassi. **SQL Planet: Jogo online para ensino de linguagem SQL**. Instituto Federal do Triângulo Mineiro Departamento de Ciência da Computação, Ituiutaba, MG, 2019.

BLINCOE, Kelly; DAMIAN, Daniela; GERMAN, Daniel M; GOUSIOS, Georgios; KALLIAMVAKOU, Eirini; SINGER, Leif. **An in-depth study of the promises and perils of mining GitHub**. Springer Science+Business Media, Nova York, 2015.

BOTTREL, Rachel do Monte. **Uma análise dos usos da lei de acesso à informação no Brasil em notícias do período de 2013 a 2015**. 2016. Dissertação (Mestrado em Ciência da Informação) – Programa de Pós-Graduação em Ciência da Informação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2016.

BRADSHAW, P. **O que é Jornalismo de Dados**. Manual de Jornalismo de Dados, 2014.

BRENOL, Marlise Viegas. **Transparência digital e jornalismo: modalidades comunicativas com uso de dados públicos**. Universidade Federal do Rio Grande do Sul, Programa de Pós-Graduação em Comunicação e Informação, Porto Alegre, 2019.

CASTELLS, Manuel. **A era da informação: economia, sociedade e cultura. In: A Sociedade em rede**. São Paulo : Paz e Terra, 2000. v. 1.

CASTELLS, Manuel. **Internet Galaxy: Reflections on the Internet, Business and Society**. Oxford Press, 2001.

CODDINGTON, Mark. **Clarifying journalism's quantitative turn: a typology for evaluating data journalism, computational journalism, and computer-assisted reporting**. Digital Journalism, 2014.

COLEMAN, G. **The political agnosticism of free and open source software and the inadvertent politics of contrast**. Anthropological Quarterly 77(3): 507–519, 2004.

CRANBERG, Lawrence. **Plea for recognition of scientific character of journalism: responsible journalists are practicing scientists**. Journalism Educator. V. 43. Columbia (EUA): Journalism& Mass Communication Educator, 1989.

CHU, Xu; ILYAS, Ihab F.; KRISHNAN, Sanjay; WANG, Jiannan. **Data Cleaning: Overview and Emerging Challenges**. SIGMOD '16: Proceedings of the 2016 International Conference on Management of Data, Páginas 2201–2206, 2016.

FONSECA, Virgínia Pradelina da Silveira et al. **Jornalismo guiado por dados como ferramenta de fact-checking: uma experiência laboratorial**. Comunicação & Inovação, v. 19, n. 41, 2018.

DADER, José Luis. **Periodismo de precisión: la vía socioinformática de descubrir noticias**. Madrid: Editorial Síntesis, 2002.

DEL VECCHIO-LIMA, Myrian Regina; SPECHT, Patrícia Pivoto. **INTERPRETAÇÕES SOBRE ALGUMAS FRAGILIDADES DO BLOG ESTADÃO DADOS**. Animus. Revista Interamericana de Comunicação Midiática, v. 20, n. 42, 2021.

DELTAFOLHA. **Material e métodos da reportagem "Questões do Enem na mira de Bolsonaro são eficientes em testar conhecimento"**, 2021. Disponível em: <https://github.com/deltafolha/enem/blob/main/metodologias/metodologia_itens_polemicos.md>. Acesso em: 09/04/2022.

DEVMOUNTAIN. **Git vs. GitHub: What's the Difference?**, |20-?|. Página "Git vs. GitHub: What's the Difference?". Disponível em <https://devmountain.com/blog/git-vs-github-whats-the-difference/>. Acesso em: 18/04/2022.

DIAKOPOULOS, Nicholas. **Cultivating the Landscape of Innovation in Computational Journalism**. CUNY Graduate School of Journalism and Tow-Knight Center for Entrepreneurial Journalism, 2012.

EXCRIPT. **Dicionário técnico de programação**, |20--?|. Página "Dicionário"**.** Disponível em: <http://excript.com/dicionario/indice.html>. Acesso em: 02/04/2022.

FIQUEM SABENDO. **Quem Somos**, |2021?|. Página "Quem Somos"**.** Disponível em: <https://fiquemsabendo.com.br/quem-somos-contato/>. Acesso em: 30/03/2022.

FLEW, T.; SPURGEON, C.; DANIEL, A.; SWIFT, A. **The promisse of computacional journalism. Journalism Practice**, 6(2):57-171. http://dx.doi.org/10.1080/17512786.2011.616655, 2012.

GAÚCHA ZH. **Práticas editoriais em GZH**, 2020. Página “Práticas Editoriais”**.** Disponível em: <https://gauchazh.clicrbs.com.br/praticas-editoriais/>. Acesso em: 30/03/2022.

GAUR, Deepti; RENU; VERMA, Tanu. **Tokenization and Filtering Process in RapidMiner**. International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA, 2014.

GEHRKE, Marília. **O resgate da objetividade como método aplicado ao jornalismo guiado por dados**. 15º ENCONTRO NACIONAL DE PESQUISADORES EM JORNALISMO. Anais... São Paulo: Universidade de São Paulo, 2017.

GEHRKE, Marília. **O uso de fontes documentais no jornalismo guiado por dados. Dissertação (Mestrado em Comunicação e Informação)**. Universidade Federal do Rio Grande do Sul. Porto Alegre: UFRGS, 2018a.

GEHRKE, Marília. **Transparência no método como valor para o jornalismo**. 16º Encontro Nacional de Pesquisadores em Jornalismo, São Paulo, SP, Brasil. Anais, v. 1, p. 1-15, 2018b.

GÊNERO E NÚMERO. **Conheça a Gênero e Número**, |20--?|. Disponível em: <https://www.generonumero.media/institucional/>. Acesso em: 30/03/2022.

GERALDES; Elen; SOUSA, Janara. **O impacto da lei de acesso à informação nas rotinas produtivas do jornalismo brasileiro**. Revista Eptic, v. 18, n. 3, p. 7-18, 2016.

GITHUB. **Give your code a home in the cloud**, 2022. Disponível em: <https://github.com/>. Acesso em: 30/03/2022.

GITHUB. **What is GitHub?**. 2016. Vídeo no canal do GitHub no YouTube. Disponível em: <https://www.youtube.com/watch?v=w3jLJU7DT5E>. Acesso em: 18/04/2022.

GOMES Jr., Paulo Pinheiro. **Big Data e o consumo de notícias nas redes sociais = Big Data and the consumption of news in social networks**. Ano 11 v. 11, n. 1, p. 46-57, jan. Revista Gestão e Desenvolvimento - Universidade Feevale, Novo Hamburgo, 2014. Metodologia (Materiais e Métodos, Hipóteses ou Questões Problemas), 2014.

GUERRA, Josenildo. **O percurso interpretativo na produção da notícia**: verdade e relevância como parâmetros de qualidade jornalística. São Cristóvão: Editora UFS; Aracaju: Fundação Oviêdo Teixeira, 2008.

GULIA, Preeti; RATRA, Ritu. **Big Data Tools and Techniques: A Roadmap for Predictive Analytics**. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-2, December, 2019.

HAMILTON, Jame. T.; TURNER, Fred. **Accountability Through Algorithm: Developing the Field of Computational Journalism**. Center For Advanced Study in the Behavioral Sciences Summer Workshop, Stanford University, 2009.

HERAVI, Bahareh. **Teaching data journalism**. In: MAIR, John et. al. (ed.). Data journalism: past, present and future. Suffolk: Abramis, 2017. p. 221-228.

INFOAMAZÔNIA. **Sobre Nós**, |20--?|. Página "Geojornalismo na Amazônia"**.** Disponível em: <https://infoamazonia.org/sobre/>. Acesso em: 30/03/2022.

KARLSSON, Michael. **Rituals of transparency: Evaluating online news outlets' uses of transparency rituals in the United States, United Kingdom and Sweden**. Journalism Studies, V. 11,N. 4. Londres: Routledge - Taylor & Francis Group, 2010. p. 535-545.

KELTY, CM. **Two Bits: The Cultural Significance of Free Software**. Durham, NC: Duke University Press, 2008.

KOVACH, Bill; ROSENSTIEL, Tom. **Os elementos do jornalismo: o que os jornalistas devem saber e o público exigir**. São Paulo: Geração Editorial, 2004.

LEONARDOJS1981. 0. **Lendo os dados e criando base para análise**. 2018. GitHub. Disponível em: <https://github.com/leonardojs1981/analise_enem_2/blob/master/02_enem_microdados_agregado_escola.ipynb>. Acesso em: 10/04/2022.

LEWIS, Seth; USHER, Nikki. **Open source and journalism: toward new frameworks for imagining news innovation.** Media, culture & society, v. 35, n. 5, p. 602-619, 2013. DOI: https://doi.org/10.1177/0163443713485494.

LIPPMANN, Walter. **Liberty and the news**. Nova York (EUA): Harcourt, Brace and Howe, 1920.

LISBOA, Silvia; BENETTI, Marcia. O jornalismo como crença verdadeira justificada. **Brazilian Journalism Research**. V. 2, N. 2. Brasília: SBPJor, 2015.

MACHADO, Elias; MaIOS, Marcos. **Um modelo híbrido de pesquisa: a metodologia aplicada pelo GJOL (2007)**. In: BARBOSA, Suzana; MACHADO, Elias; PALACIOS, Marcos (Org.). GJOL 20 anos de percurso. Textos fundadores e metodológicos. Salvador: EDUFBA, 2018. p. 341-363.

MANSELL, Robin, WEHN, Uta. **Knowledge societies: information technologies for sustainable development**. Oxford : Oxford University, 1998.

MANCINI, Leonardo; VASCONCELLOS, Fabio. **Jornalismo de Dados: conceito e categorias**. Fronteiras-estudos midiáticos, v. 18, n. 1, p. 69-82, 2016.

MARIANI, Daniel et. al. **Estudo inédito indica alta chance de fraude em mil provas do Enem**. São Paulo: Folha de S.Paulo, 2018. Disponível em: <https://www1.folha.uol.com.br/educacao/2018/04/estudo-inedito-indica-alta-chance-de-fraude-em-mil-provas-do-enem.shtml>. Acesso em 19/04/2022.

MENEZES, Nilo Ney Coutinho. I**ntrodução à programação com Python**. Novatec Editora Ltda., São Paulo, SP, Brasil, 2010.

MEYER, Philip. **Precision Journalism: A Reporter's Introduction to Social Science Methods**. Indiana, EUA: Indiana University Press, 1973.

MEYER, Philip. **Precision Journalism**. A Reporter´s Introduction to Social Science Methods. New York: Rowman & Littlefield Publishers, 2002.

MEYER, Philip. **The next journalism's objective reporting**. Nieman Reports, n. 58, 2004.

MIELNICZUK, Luciana. **Jornalismo na Web: uma contribuição para o estudo do formato da notícia na escrita hipertextual**. Tese (Doutorado em Comunicação e Cultura Contemporâneas), UFBA, Salvador, 2003.

NEXO JORNAL. **Sobre o NEXO**, 2022. Página "Sobre o NEXO"**.** Disponível em: <https://www.nexojornal.com.br/sobre/Sobre-o-Nexo>. Acesso em: 30/03/2022.

NUNES, Greyce Ellen Vargas. **Os efeitos da audiência digital e a busca por inovação nas redações de GaúchaZH e Folha de S. Paulo**. 2018. 144 páginas. Programa de Pós-Graduação em Comunicação - Universidade do Vale do Rio dos Sinos, São Leopoldo, 2018.

OPEN DEFINITION. **The Open Definition**. [Online] 2015. Disponível em: <http://opendefinition.org/>.

OPEN KNOWLEDGE FOUNDATION. **Open Data Handbook**. 2010. Disponível em: <http://opendatahandbook.org/guide/en/>.

PERKEL, Jeffrey. **When it comes to reproducible science, Git is code for success**. Nature Index, 2018. Disponível em <https://www.natureindex.com/news-blog/when-it-comes-to-reproducible-science-git-is-code-for-success>. Acesso em: 17/04/2022.

PHILLIPS, A. **Transparency and the new ethics of journalism**. Journalism Practice 4(3): 373–382, 2010.

PINHEIRO, Paulo César da Costa. **Desenvolvimento de um tutorial hipertexto em HTML(Hyper Text Markup Language)**. Anais da XXV Cobenge. Escola Politécnica da UFBA, Salvador, 1997. Disponível em <https://www.researchgate.net/profile/Paulo-Cesar-Pinheiro/publication/334971400_DESENVOLVIMENTO_DE_UM_TUTORIAL_HIPERTEXTO_EM_HTML/links/5d4846

ee92851cd046a353dd/DESENVOLVIMENTO-DE-UM-TUTORIAL-HIPERTEXTO-EM-HTML.pdf>. Acesso em 02/04/2022.

POVOA, Lucas Venezian; MANZIONE, Rodrigo Lilla; WENDLAND, Edson Cesar. **Rotinas para Análises Geoestatísticas Utilizando a Linguagem R: um exemplo com dados agro-ambientais**. Simpósio de Geoestatística aplicada em ciências agrárias, 2011.

RAUSCHMAYER, A. **Speaking JavaScript: An In-Depth Guide for Programmers**. 1. ed. Califórnia: O'Reilly Media, 2014.

RAVAL, Kalyani M. **Data Mining Techniques**. International Journal of Advanced Research in Computer Science and Software Engineering, 2012.

REFERÊNCIAS NO BRASIL. **Dados Abertos Pernambuco**, 2020. Disponível em: <https://www.dadosabertospernambuco.com.br/jornalismodedadosbr>. Acesso em: 30/03/2022.

RIEHLE, Dirk. **Framework Design: A Role Modeling Approach**. Ph.D. Thesis, No. 13509. Zürich, Switzerland, ETH Zürich, 2000.

ROBINSON, S. '**Journalism as process': the labor implications of participatory content in news organization**. Journalism & Communication Monographs, 2011. 13(3): 138–210.

SALAVERRÍA R.; GARCÍA AVILÉS. J.A.; MASIP P.M. **Concepto de Convergencia Periodística**. In: LÓPEZ GARCÍA, X.; PEREIRA FARIÑA, X. Convergencia Digital. Reconfiguração de los Medios de Comunicación en España. Santiago de Compostela: Universidade de Santiago de Compostela, 2010.

SIRISURIYA, S.C.M. de S. **A Comparative Study on Web Scraping**. Department of Computer Science, Faculty of Computing, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka, 2015.

STAVELIN, Erik. **Computational journalism: when journalism meets programming**. Thesis. Department of Information Science and Media Studies, University of Bergen, Norway, 2013.

STRAY, J. 2014. **The Data Journalist´s Eyes, An Introduction**. Disponível em <https://medium.com/tow-center/the-data-journalists-eye-an-introduction-7ac8f24a4099> . Acesso em: 19/04/2022.

SU, Jiang; ZHANG, Harry. **A Fast Decision Tree Learning Algorithm**. Faculty of Computer Science University of New Brunswick, NB, Canada, E3B 5A3, 2006.

THE WHITE HOUSE - PRESIDENT BARACK OBAMA. **About Open Government**, |201?|. Página "Open Government Iniative". Disponível em: <https://obamawhitehouse.archives.gov/open/about>. Acesso em: 16/04/2022.

TRÄSEL, Marcelo. **Jornalismo guiado por dados: relações da cultura hacker com a cultura jornalística**. In: Encontro Anual da Compós, XXII, Bahia. Anais... Universidade Federal da Bahia, 2013. Disponível em: <http://www.academia.edu/3136931/JORNALISMO_GUIADO_POR_DADOS_rela%C3%A7%C3%B5es_da_cultura_hacker_com_a_cultura_jornal%C3%ADstica>. Acesso em: 19/04/2022.

TRÄSEL, Marcelo. **Entrevistando planilhas: estudo das crenças e do ethos de um grupo de profissionais de jornalismo guiado por dados no Brasil**. Tese (Doutorado em Comunicação Social), PUCRS, Porto Alegre, 2014.

TRÄSEL, Marcelo. **Hacks and hackers: the ethos and beliefs of a group of Data-Driven Journalism professionals in Brazil**. Revista FAMECOS, 25(1), ID27589, 2018.

TSAY, J.; DABBISH, L.; HERBSLEB, J. **Influence of social and technical factors for evaluating contribution in GitHub**. In (pp. 356–366). NewYork, NY:ACM, 2014.

TURNER, F. **Where the counterculture met the new economy: the WELL and the origins of virtual community**. Technology and Culture, 2005. 46(3): 485–512.

UC SANTA CATARINA. **If-Then Statements**, 2016. Página “Data Management Services”. Disponível em: <https://datamgmt.ucsc.edu/help-training/infoview-help/how-do-i/variables/formula-basics/if-then.html>. Acesso em: 11/04/2022.

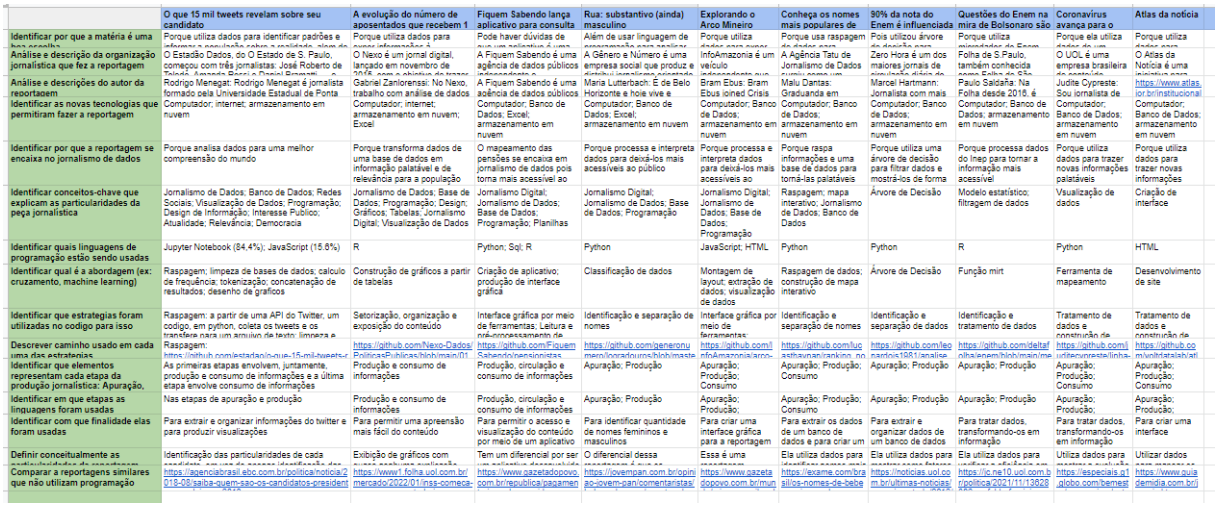
UOL. **O Grupo UOL é a maior empresa brasileira de conteúdo, tecnologia, serviços e meios de pagamentos**, 2021. Página “História”. Disponível em: <https://sobreuol.noticias.uol.com.br/historia/>. Acesso em: 30/03/2022.

WARD, Stephen J. A. **Ethics and the Media**. Cambridge: Cambridge University Press, 2011.

WEINBERGER, David. **Transparency: the new objectivity**. In: Tred-Setting Products, V. 18. Camden: KM World, 2009. Disponível em: <http://www.kmworld.com/Articles/Column/David-Weinberger/Transparency-the-new-objectivity-55785.aspx>. Acesso em: 19/04/2022.

WERTHEIN, Jorge. **A sociedade da informação e seus desafios**. Ci. Inf., Brasília, v. 29, n. 2, 2000. Disponível em <https://doi.org/10.1590/S0100-19652000000200009>. Acesso em: 06/04/2022.

APÊNDICE A - PLANILHA DE ANÁLISE DO CORPUS - PERGUNTAS



	O que 15 mil tweets revelam sobre seu candidato	A evolução do número de aposentados que recebem 1	Fiquem Sabendo lança aplicativo para consulta	Rua: substantivo (ainda) masculino	Explorando o Arco Mineiro	Conheça os nomes mais populares de	90% da nota do Enem é influenciada	Questões do Enem na mira de Bolsonaro são	Coronavírus avança para o	Atlas da notícia
Identificar por que a matéria é uma	Porque utiliza dados para identificar padrões e	Porque utiliza dados para	Pode haver dúvidas de	Além de usar linguagem de	Porque utiliza	Porque usa raspagem	Pois utilizou árvore	Porque utiliza	Porque ela utiliza	Porque utiliza
Análise e descrição da organização jornalística que fez a reportagem	O Estadão Dados, do O Estado de S. Paulo,	O Nexo é um jornal digital,	A Fiquem Sabendo é uma	A Gênero e Número é uma	InfoAmazonia é um	A Agência Tatu de	Zero Hora é um dos	Folha de S.Paulo,	O UOL é uma	O Atlas da
Análise e descrições do autor da reportagem	Rodrigo Menegat; Rodrigo Menegat é jornalista	Gabriel Zanlorenssi; No Nexo,	A Fiquem Sabendo é uma	Maria Lutterbach: É de Belo	Bram Ebus: Bram	Malu Dantas:	Marcel Hartmann:	Paulo Saldaña: Na	Judite Cypreste:	https://www.atlas.
Identificar as novas tecnologias que permitiram fazer a reportagem	Computador; internet; armazenamento em	Computador; internet;	Computador; Banco de	Computador; Banco de	Computador; Banco	Graduanda em	Jornalista com mais	Folha desde 2016, é	Sou jornalista de	jor.br/institucional
	nuvem	armazenamento em nuvem;	Dados; Excel;	Dados; Excel;	de Dados;	Computador; Banco	Computador; Banco	Computador; Banco de	Computador;	Computador;
		Excel	armazenamento em	armazenamento em nuvem	armazenamento em	de Dados;	de Dados;	Dados; armazenamento	Banco de Dados;	Banco de Dados;
			nuvem		nuvem	armazenamento em	armazenamento em	em nuvem	armazenamento	armazenamento
						nuvem	nuvem		em nuvem	em nuvem
Identificar por que a reportagem se encaixa no jornalismo de dados	Porque analisa dados para uma melhor	Porque transforma dados de	O mapeamento das	Porque processa e interpreta	Porque processa e	Porque raspa	Porque utiliza uma	Porque processa dados	Porque utiliza	Porque utiliza
	compreensão do mundo	uma base de dados em	pensões se encaixa em	dados para deixá-los mais	interpreta dados	informações e uma	árvore de decisão	do Inep para tornar a	dados para trazer	dados para
		informação palatável e de	jornalismo de dados pois	acessíveis ao público	para deixá-los mais	base de dados para	para filtrar dados e	informação mais	novas informações	trazer novas
		relevância para a população	torna mais acessível ao		acessíveis ao	torná-las palatáveis	mostrá-los de forma	acessível	palatáveis	informações
					público					
Identificar conceitos-chave que explicam as particularidades da peça jornalística	Jornalismo de Dados; Banco de Dados; Redes Sociais; Visualização de Dados; Programação; Design de Informação; Interesse Publico; Atualidade; Relevância; Democracia	Jornalismo de Dados; Base de Dados; Programação; Design; Gráficos; Tabelas; Jornalismo Digital; Visualização de Dados	Jornalismo Digital; Jornalismo de Dados; Base de Dados; Programação; Planilhas	Jornalismo Digital; Jornalismo de Dados; Base de Dados; Programação	Jornalismo Digital; Jornalismo de Dados; Base de Dados; Programação	Raspagem; mapa interativo; Jornalismo de Dados; Banco de Dados	Árvore de Decisão	Modelo estatístico; filtragem de dados	Visualização de dados	Criação de interface
Identificar quais linguagens de programação estão sendo usadas	Jupyter Notebook (84.4%); JavaScript (15.6%)	R	Python; Sql; R	Python	JavaScript; HTML	Python	Python	R	Python	HTML
Identificar qual é a abordagem (ex: cruzamento, machine learning)	Raspagem; limpeza de bases de dados; calculo de frequência; tokenização; concatenação de resultados; desenho de graficos	Construção de gráficos a partir de tabelas	Criação de aplicativo; produção de interface gráfica	Classificação de dados	Montagem de layout; extração de dados; visualização de dados	Raspagem de dados; construção de mapa interativo	Árvore de Decisão	Função mirt	Ferramenta de mapeamento	Desenvolvimento de site
Identificar que estrategias foram utilizadas no codigo para isso	Raspagem: a partir de uma API do Twitter, um codigo, em python, coleta os tweets e os transfere para um arquivo de texto; limpeza e	Setorização, organização e exposição do conteúdo	Interface gráfica por meio de ferramentas; Leitura e pré-processamento de	Identificação e separação de nomes	Interface gráfica por meio de ferramentas;	Identificação e separação de nomes	Identificação e separação de dados	Identificação e tratamento de dados	Tratamento de dados e construção de	Tratamento de dados e construção de
Descrever caminho usado em cada uma das estrategias	Raspagem:	https://github.com/Nexo-Dados/	https://github.com/Fiquem	https://github.com/generonu	https://github.com/l	https://github.com/luc	https://github.com/del	https://github.com/ol	https://github.co	https://github.co
	https://github.com/estadao/o-que-15-mil-tweets-r	PoliticasPublicas/blob/main/01	Sabendo/pensionistas	mero/logradouros/blob/maste	nfoAmazonia/arqo-	asthayanan/ranking_no	nardos1981/analise-	olha/enem/blob/main/me	uideucpreste/olha-	m/politatalah/ati
Identificar que elementos representam cada etapa da produção jornalística: Apuração,	As primeiras etapas envolvem, juntamente, produção e consumo de informações e a última etapa envolve consumo de informações	Produção e consumo de informações	Produção, circulação e consumo de informações	Apuração; Produção	Apuração; Produção; Consumo	Apuração; Produção; Consumo	Apuração; Produção	Apuração; Produção	Apuração; Produção; Consumo	Apuração; Produção; Consumo
Identificar em que etapas as linguagens foram usadas	Nas etapas de apuração e produção	Produção e consumo de informações	Produção, circulação e consumo de informações	Apuração; Produção	Apuração; Produção;	Apuração; Produção; Consumo	Apuração; Produção	Apuração; Produção	Apuração; Produção;	Apuração; Produção;
Identificar com que finalidade elas foram usadas	Para extrair e organizar informações do twitter e para produzir visualizações	Para permitir uma apreensão mais fácil do conteúdo	Para permitir o acesso e visualização do conteúdo por meio de um aplicativo	Para identificar quantidade de nomes femininos e masculinos	Para criar uma interface gráfica para a reportagem	Para extrair os dados de um banco de dados e para criar um	Para extrair e organizar dados de um banco de dados	Para tratar dados, transformando-os em informação	Para tratar dados, transformando-os em informação	Para criar uma interface
Definir conceitualmente as particularidades da reportagem	Identificação das particularidades de cada	Exibição de gráficos com	Tem um diferencial por ser	O diferencial dessa	Essa é uma	Ela utiliza dados para	Ela utiliza dados para	Ela utiliza dados para	Utiliza dados para	Utilizar dados
Comparar a reportagens similares que não utilizam programação	https://agenciabrasil.ebc.com.br/politica/noticia/2 018-08/saiba-quem-sao-os-candidatos-president	https://www1.folha.uol.com.br/ mercado/2022/01/inss-comeca-	https://www.gazetadopovo. com.br/republica/pagamen	https://overman.com.br/opini ao-jovem-pan/comentaristas/	https://www.gazeta dopovo.com.br/mun	https://exame.com/bra sil/os-nomes-de-bebe	https://noticias.uol.co m.br/ultimas-noticias/	https://c.ne10.uol.com.b r/politica/2021/11/13828	https://especiais.g1 .globo.com/bemest	https://www.qua demidia.com.br/j

Link para visualização:

https://docs.google.com/spreadsheets/d/1Zp6dM1xczzJh98Ts_InMnGxPUBwybY_trPaDhzIEkfs/edit?usp=sharing

APÊNDICE B - PLANILHA DE ANÁLISE DO CORPUS - CLASSIFICAÇÃO

Título	Veículo	Tamanho do veículo	Tema	Data de Publicação	Autores	Tipos de fontes	Base de dados utilizada	Pública ou outro tipo?	Objetivo	Linguagens de programação utilizadas	Processos utilizados
Conheça os nomes mais populares de 2021 em cada estado brasileiro	Agência Tatu	Médio	Cotidiano	22/12/2021	Malu Dantas; Lucas Thaynan	Apenas documentais (banco de dados público)	Portal Oficial de Registro Civil	Pública	Mostrar os nomes mais comuns do ano, no geral e por estado	Python	Raspagem; visualização interativa de dados
Fiquem Sabendo lança aplicativo para consulta de remuneração de pensionistas	Fiquem Sabendo	Pequeno	Política	28/07/2021	Fiquem Sabendo	Apenas documentais (banco de dados público)	Ministério da Defesa / Portal de Transparência	Pública	Mostrar remuneração de pensionistas	Python; Sql; R	Criação de aplicativo; produção de interface gráfica
Questões do Enem na mira de Bolsonaro são eficientes em testar conhecimento	Folha de São Paulo	Grande	Educação; Política	20/11/2021	Paulo Saldaña; Daniel Mariani; Diana Yukari; Flávia Faria	Documentais; Especializadas; Oficiais	Inep	Pública	Identificar se as questões criticadas por Bolsonaro são eficientes na assimilação dos conteúdos	R	Modelo Estatístico
90% da nota do Enem é influenciada por fatores econômicos e culturais, indica análise feita por GaúchaZH com uso de inteligência artificial	Gaúcha ZH	Grande	Educação	28/07/2019	Marcel Hartmann; Hermes Wiederkehr; Anna Fernandes	Documentais; Especializadas; Não oficiais	Inep	Pública	Identificar como condições sociais influenciam na nota do Enem	Python	Árvore de Decisões
Rua: substantivo (ainda) masculino	Gênero e Número	Pequeno	Igualdade	15/03/2017	Maria Lutterbach; Álvaro Justen; Natália Mazotte	Apenas documentais (banco de dados privado)	Empresa de Correios e Telegrafos (ECT)	Privada	Descobrir quantos logradouros têm nomes de homens e mulheres	Python	Classificação de dados
Explorando o Arco Mineiro	InfoAmazonia	Pequeno	Direitos Humanos; Meio Ambiente	15/01/2018	Bram Ebus	Documentais; Oficiais; Não Oficiais	Inflação FMI; Cesta de referência do petróleo OPEC; Malária OMS	Privada e Pública	Mostrar dados sobre inflação, preço do petróleo e malária para explicar situação da Venezuela	JavaScript; HTML	Montagem de layout; extração de dados; visualização de dados
A evolução do número de aposentados que recebem 1 salário mínimo	Nexo	Médio	Economia	22/07/2020	Gabriel Zanlorenssi; Lucas Gomes	Apenas documentais (banco de dados público)	INSS	Pública	Descobrir a evolução do número de aposentados que recebe 1 salário mínimo	R	Produção de visualização de dados
O que 15 mil tweets revelam sobre seu candidato	O Estado de São Paulo	Grande	Política	09/05/2018	Rodrigo Menegat; Cecília do Lago; Vinicius Sueiro; Carlos Marin	Apenas documentais (tweets)	Twitter	Privada	Descobrir tweets mais característicos de cada candidato	Python; Javascript	Raspagem; limpeza de bases de dados; calculo de frequência; tokenização; concatenação de resultados; desenho de graficos
Coronavírus avança para o interior do Brasil; veja evolução em mapa	UOL	Grande	Saúde	16/04/2020	Judite Cypreste	Apenas Documentais (banco de dados privado e público)	Brasil.io; Fiocruz	Privada e Pública	Mostrar a evolução temporal do coronavírus no Brasil	Python	Visualização de dados
Atlas da notícia	Volt Data Lab	Pequeno	Informação	01/11/2017	https://www.atlas.jor.br/institucional/equipe-do-atlas-da-noticia/	Documentais (dados públicos); contribuições voluntárias	Secom	Pública	Mapear o jornalismo no Brasil	HTML	Exposição de dados

Link para visualização:

https://docs.google.com/spreadsheets/d/1Zp6dM1xczzJh98Ts_InMnGxPUBwybY_trPaDhzIEkfs/edit?usp=sharing