Jornalismo de Dados: transformação digital na produção de notícias

Henrique Bilo[1], **Rafael Oleques Nunes**[1], **Daniel Matos de Castro**[1], **Dante Augusto Couto Barone**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brasil

{henrique.bilo, dmcastro, ronunes, barone}@inf.ufrgs.br

Abstract. *This article aims to explore the use of data journalism as an emerging technique in the field of modern journalism. To do so, we present concrete examples of articles and newspapers that use this technique to collect, analyze, and visualize large sets of data. We highlight how data journalism allows for more accurate and factual reporting, while identifying patterns and trends in large amounts of information. We also emphasize the challenges that digital transformation presents for journalism, such as combating fake news, as well as opportunities for the production of more diverse and secure content. Finally, we conclude that this is a promising and essential area for modern journalism, contributing to greater accuracy and transparency in the information disseminated to society.*

Resumo. *Este artigo tem como objetivo explorar o uso do jornalismo de dados como uma técnica emergente no campo do jornalismo moderno. Para isso, apresentamos exemplos concretos de matérias e jornais que utilizam essa técnica para coletar, analisar e visualizar grandes conjuntos de dados. Destacamos como o jornalismo de dados permite produzir reportagens mais precisas e factuais, ao mesmo tempo em que identifica padrões e tendências em grandes quantidades de informações. Enfatizamos, também, os desafios que a transformação digital apresenta para o jornalismo, como o combate às fake news, bem como as oportunidades para a produção de conteúdo mais diverso e seguro. Por fim, concluímos que trata-se de uma área promissora e essencial para o jornalismo moderno, contribuindo para uma maior precisão e transparência nas informações divulgadas para a sociedade.*

1. Introdução

Nos últimos anos, com o aumento da quantidade de dados disponíveis e da necessidade de torná-los acessíveis para a tomada de decisões, o jornalismo de dados se tornou uma ferramenta indispensável para jornalistas e veículos de comunicação [Bounegru and Gray 2021]. Com a ajuda de técnicas avançadas de análise e visualização de dados, é possível identificar tendências e padrões complexos, que muitas vezes passariam despercebidos em análises tradicionais. Além disso, o jornalismo de dados tem se mostrado uma importante arma na luta contra as chamadas *fake news*, ao permitir a verificação de informações e a checagem de dados divulgados . Com o uso de fontes confiáveis e técnicas de verificação, os jornalistas podem desmistificar informações falsas que circulam na internet e na sociedade.

Porém, como em todas as áreas do jornalismo, o jornalismo de dados também enfrenta desafios, como a falta de acesso a dados governamentais [Fink and Anderson 2015], a dificuldade em encontrar profissionais capacitados [Aitamurto et al. 2011] e a necessidade de se manter atualizado com as novas tecnologias e ferramentas de análise de dados.

Neste contexto, é fundamental discutir e refletir sobre o papel do jornalismo de dados na sociedade atual e entender como essa técnica pode ser usada de forma ética e responsável na produção de conteúdo jornalístico. Busca-se, portanto, fornecer uma análise crítica sobre a aplicação dessa tecnologia na sociedade, considerando seus possíveis benefícios.

Na seção 2, apresentamos o conceito de jornalismo de dados e sua utilidade na produção de reportagens mais precisas e contextualizadas. Já na seção 3, exploramos o potencial da inteligência artificial no jornalismo, exemplificando com ferramentas como o ChatGPT. Além disso, na seção 4 abordamos um dos principais desafios dessa área: o combate às notícias falsas. Por último, na seção 5, fazemos algumas considerações finais sobre a importância do jornalismo de dados para a prática jornalística em geral.

2. Jornalismo de dados

O jornalismo de dados é uma área do jornalismo que utiliza ferramentas digitais e análises estatísticas para coletar, organizar e interpretar grandes volumes de informações. Desde sua primeira aparição em 1821 [Gray et al. 2012], essa prática tem se mostrado cada vez mais relevante para a sociedade, permitindo a produção de reportagens mais profundas, precisas e contextualizadas.

Com o avanço da tecnologia digital, o jornalismo de dados tornou-se mais sofisticado, permitindo aos jornalistas utilizar ferramentas de visualização de dados, inteligência artificial e aprendizado de máquina para analisar grandes conjuntos de dados e produzir reportagens complexas. Isso se deve, em parte, ao crescente volume de dados virtuais disponíveis, já que cada vez mais pessoas utilizam a internet para se manter informadas.

Além disso, o jornalismo de dados é uma ferramenta poderosa para investigar casos de corrupção, má gestão e outras formas de abuso de poder. Com o uso de dados, é possível documentar fatos e construir evidências para apresentar ao público e às autoridades competentes. Vale destacar que o jornalismo de dados não se limita a fontes de dados públicas, podendo incluir dados obtidos por meio de investigação independente ou fontes confidenciais.

Por isso, é fundamental que os jornalistas deem cada vez mais importância ao jornalismo de dados e aprimorem suas habilidades em programação, estatística, visualização de dados e outras áreas técnicas. Dessa forma, poderão produzir reportagens mais atraentes e confiáveis, que ofereçam uma análise precisa e profunda dos fatos para o público.

2.1. Recursos e Ferramentas

Nesta seção, exploraremos os recursos e ferramentas essenciais para o jornalismo de dados. O jornalismo de dados é uma área crescente que exige habilidades técnicas e analíticas, bem como a capacidade de comunicar informações complexas de forma clara

e acessível. Para ajudar os jornalistas a realizar análises de dados e produzir reportagens de qualidade, existem diversas ferramentas e recursos disponíveis. Vamos explorar algumas dessas ferramentas e recursos importantes.

- *Big Data*: As ferramentas de *Big Data* ajudam os jornalistas a organizar e analisar grandes quantidades de informações para produzir reportagens com informações mais ricas, científicas e democráticas [Hammond 2017]. Com elas, é possível filtrar e encontrar dados importantes de forma mais fácil. Além disso, são utilizados *dashboards* que tornam a análise de dados mais simples e visualmente atraente. Com isso, é possível criar reportagens com imagens, vídeos e infográficos que prendem a atenção do leitor. Como exemplos dessas ferramentas, podemos citar:
  - Apache Spark [1]: Plataforma de processamento de dados em cluster que é utilizada para analisar e processar grandes quantidades de dados;
  - Apache Cassandra [2]: Banco de dados capaz de lidar com grandes quantidades de dados em tempo real;
  - Tableau [3]: Permite a criação de gráficos interativos, tabelas e outros elementos visuais para apresentar informações de maneira atraente e clara;
  - Python [4]: Linguagem de programação frequentemente utilizada para coletar e analisar dados que utiliza bibliotecas como *Pandas* [5] e *scikit-learn* [6];
- Análise Semântica: As ferramentas de análise semântica coletam palavras-chave da internet que representam as buscas de informação em motores de busca e comentários em redes sociais ou fóruns. Com a ajuda de aplicativos de inteligência semântica, os jornalistas podem entender o que os usuários procuram e quais temas são mais relevantes para eles [Broussard et al. 2019], gerando ideias para conteúdo com maior personalização. Entre esses aplicativos, podemos citar:
  - Google Trends [7]: Permite visualizar as tendências de busca dos usuários em todo o mundo. É possível visualizar quais são as palavras-chave mais populares em um determinado período e em uma determinada região.
  - SEMrush [8]: Parecido com o Google Trends, essa ferramenta oferece recursos para análise de palavras-chave, análise de concorrência e análise de tráfego de sites. Com isso, o usuário consegue entender quais palavras-chave são as mais relevantes em um determinado nicho ou segmento.
  - BuzzSumo [9]: Permite ao usuário descobrir o conteúdo mais popular em um determinado nicho ou segmento. Dessa forma, o usuário pode encontrar tópicos de interesse para o público-alvo e identificar oportunidades de novas reportagens.

[1] https://spark.apache.org
[2] https://cassandra.apache.org
[3] https://www.tableau.com/pt-br
[4] https://www.python.org
[5] https://pandas.pydata.org/
[6] https://scikit-learn.org/stable/
[7] https://trends.google.com.br/home
[8] https://pt.semrush.com
[9] https://buzzsumo.com

- Github [10]: Plataforma utilizada por programadores para compartilhar ideias e criar *softwares* de código aberto de forma que as empresas de notícias podem usar para criar seus próprios aplicativos.

2.1.1. Visualizações

Com a crescente quantidade de dados complexos, informações abstratas e de métricas que não são lidadas no cotidiano, torna-se relevante o uso de visualizações para facilitar a compreensão dos significados e das relações entre os dados. Essa compreensão tende a ser mais eficaz por meio de características visuais [Bounegru and Gray 2021]. As propriedades perceptivas das representações gráficas são especialmente úteis para visualizar e entender esses dados [Bounegru and Gray 2021], o que tem levado ao desenvolvimento de novas ferramentas e técnicas tanto na academia quanto na indústria.

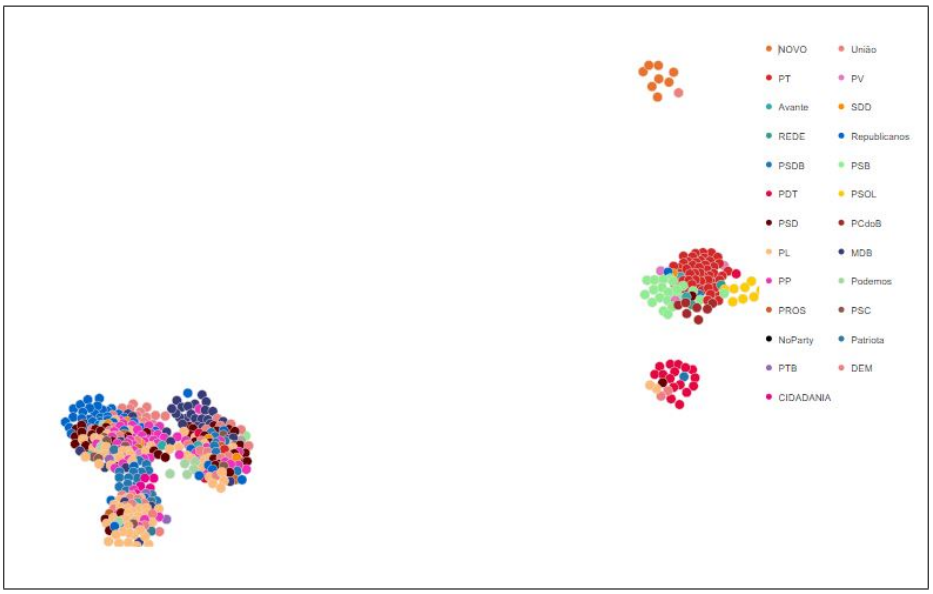

Figure 1. Exemplo de visualização com o espectro político de deputados durante a legislatura. Os deputados que votam de forma mais semelhante tendem a ficar mais próximos, permitindo a análise de sua posição em relação ao partido (cor do deputado) e a identificação de grupos com ideologias semelhantes. Screenshot retirado da ferramenta disponível no link: https://www.inf.ufrgs.br/∼rnmsilva/CivisAnalysis2/.

A utilização de visualizações no jornalismo de dados traz diversos benefícios, incluindo a possibilidade de apresentar ao público múltiplas perspectivas sobre um mesmo assunto. Isso permite que o leitor tenha uma compreensão mais completa dos fatos e possa descobrir abordagens diferentes que possam despertar o seu interesse [Aitamurto et al. 2011].

Uma pesquisa recente realizada [da Silva et al. 2018] tem como foco auxiliar a descoberta de narrativas e perfis políticos por meio de visualizações de dados. Essa pesquisa apresenta ferramentas que permitem analisar o espectro dos deputados com base em suas votações na Câmara dos Deputados do Brasil, como mostra a Figura 1. Essa ferramenta

[10]https://github.com

permite que o público tenha uma visão mais completa da política e de seus representantes, identificando as narrativas e nuances que estão por trás dos fatos noticiados e da fala de seus candidatos. Essa abordagem de análise de dados pode ajudar o público a entender melhor o funcionamento da política e a tomar decisões mais conscientes.

Apesar de visualizações gráficas serem amplamente utilizadas, nem sempre elas são adequadas para alguns tipos de dados, podendo trazer excessos de informação ou confundindo o leitor [Aitamurto et al. 2011]. Em alguns casos, tabelas podem ser mais interessantes, tanto por se ter um conjunto menor de dados como por terem uma estrutura mais simples e direta. Outro exemplo é na utilização de valores numéricos com uma maior precisão, algo que em muitas visualizações, como em gráficos de barras, podem não ser claras. Dessa forma, destaca-se a importância de realizar uma análise prévia dos tipos de dados e qual a melhor forma de passar a informação desejada ao leitor.

2.2. Exemplos de matérias

O caso Pandora Papers [11] é um exemplo recente de como o jornalismo de dados pode revelar informações cruciais sobre o comportamento de indivíduos e empresas ao redor do mundo. Trata-se de uma investigação jornalística internacional realizada em 2021 por um consórcio de jornalistas de mais de 100 veículos de mídia em todo o mundo. Os jornalistas tiveram acesso 11.9 milhões de registros de empresas offshore e contas bancárias secretas que eram usadas por pessoas e empresas de alto perfil para esconder ativos, evitar impostos e, em alguns casos, ocultar atividades criminosas. Os documentos eram em sua grande maioria não-estruturados, incluindo e-mails, contratos e registros bancários nos mais diversos formatos de arquivo, incluindo mais de 4 milhões de arquivos PDF e quase 2 milhões de documentos do Microsoft Word. Para minerar os dados relevantes em meio à grande quantidade de informações obtidas, jornalistas utilizaram uma variedade de métodos de análise de dados, incluindo análise de rede, mineração de texto e análise de dados financeiros e algoritmos de aprendizado de máquina para extrair informações relevantes.

Outro projeto relevante foi realizado pelo The Guardian sob o nome de “The Counted” [12]. Os jornalistas analisaram dados sobre mortes causadas pela polícia nos Estados Unidos através de uma combinação de técnicas de análise de dados, incluindo limpeza de dados, categorização e análise estatística, para identificar padrões e tendências nos dados. Ao fazer isto, o The Guardian conseguiu descobrir um número desproporcional de afro-americanos mortos pela polícia, o que desencadeou um debate público nacional sobre reforma policial. Para tornar os dados acessíveis aos leitores e incentivar mais exploração, o jornal utilizou visualizações de dados e ferramentas interativas. Essas ferramentas incluíram um mapa de mortes causadas pela polícia, uma linha do tempo de eventos e uma ferramenta para explorar os dados por fatores demográficos e geográficos.

Por último, a matéria intitulada ”The Tennis Racket” [13] foi uma investigação conduzida pela BuzzFeed News e pela BBC que consistiu no uso de diversos métodos de análise de dados para descobrir evidências de uma ampla manipulação de resultados no

[11] https://www.computerweekly.com/news/252510313/Pandora-Papers-How-journalists-mined-terabytes-of-offshore-data-to-expose-the-worlds-elites

[12] https://www.theguardian.com/us-news/2015/dec/31/the-counted-police-killings-2015-young-black-men

[13] https://www.buzzfeednews.com/article/heidiblake/the-tennis-racket

tênis profissional. A investigação envolveu a análise de dados de mais de 26.000 partidas, incluindo probabilidades de apostas, pontuações e estatísticas de desempenho dos jogadores. Os resultados da investigação foram apresentados usando visualizações interativas de dados, incluindo mapas de calor e gráficos, que ajudaram a ilustrar a escala e a complexidade do problema. Ao empregar métodos rigorosos de análise de dados, a investigação foi capaz de descobrir um problema generalizado no tênis profissional, levando a discussões de medidas destinadas a prevenir futuros escândalos de manipulação de resultados.

3. Jornalismo e inteligência artificial (IA)

A inteligência artificial (IA) é outra tecnologia com um potencial de impacto significativo no jornalismo [Vicente and Flores 2021], pois pode transformar a maneira como as notícias são criadas, distribuídas e analisadas. Uma maneira pela qual a IA está influenciando o jornalismo é através da personalização. Técnicas como clusterização e reconhecimento de entidades nomeadas podem ser empregadas para identificar grupos de leitores e seus interesses, fornecendo *insights* valiosos sobre o engajamento dos leitores com o conteúdo que podem levar a definição de pautas personalizadas [Broussard et al. 2019]. Com a capacidade de analisar o comportamento dos leitores e suas preferências, a IA pode personalizar a entrega de notícias, criando *feeds* personalizados de notícias e alertas adaptados aos interesses individuais.

Porém, a chamada inteligência artificial generativa representa uma poderosa força disruptiva no jornalismo. Modelos de IA generativa podem gerar texto semelhante ao humano com base em um *prompt*, tornando-se uma ferramenta poderosa para a criação automatizada de conteúdo de alta qualidade [Pavlik 2023]. O ChatGPT, modelo de linguagem treinado pela OpenAI, com base na arquitetura GPT-3 [Wang et al. 2023], é um exemplo de modelo que pode ajudar os jornalistas em seu trabalho. Este pode gerar resumos, artigos e até mesmo relatórios inteiros usando algoritmos de geração de linguagem natural (NLG).

Um dos benefícios da IA generativa no jornalismo é a capacidade de criar rapidamente histórias de notícias a partir de grandes conjuntos de dados ou informações complexas [Pavlik 2023]. Por exemplo, o ChatGPT pode gerar um artigo de notícias sobre um relatório financeiro complexo ou um estudo científico, permitindo que os jornalistas relatem desenvolvimentos importantes rapidamente e com precisão.

No entanto, o uso da IA generativa no jornalismo não está isento de desafios. Uma das principais preocupações é o potencial de viés nos dados usados para treinar modelos de IA [Pavlik 2023]. Se os dados usados para treinar um modelo de IA generativa forem tendenciosos, a saída também pode ser tendenciosa, o que poderia ter sérias consequências para a precisão e justiça da reportagem de notícias.

Outro problema é a chamada "alucinação": com alguma frequência modelos generativos geram texto que parece coerente e fluente, mas que na verdade é factualmente incorreto ou até mesmo sem sentido [Bang et al. 2023]. No jornalismo, a precisão e credibilidade da informação são de extrema importância. Portanto, o risco de publicar uma "alucinação" pode ser particularmente prejudicial. Por exemplo, se o ChatGPT gerar um artigo de notícias contendo informações imprecisas ou enganosas, pode erodir a confiança do público na mídia. Para mitigar este problema, é essencial que os jornalistas usem ferra-

mentas de IA generativa com responsabilidade e cautela. Isso inclui tomar medidas para verificar a precisão das informações geradas pelos modelos e evitar seu uso para gerar artigos sem supervisão e edição adequadas.

4. Desafios em jornalismo de dados: Fake News

Segundo [Canavilhas and Jorge 2022] as *fake news*, com o aumento da digitalização, criaram uma "guerra de falsificação" onde as pessoas não conseguem identificar se as notícias são reais ou não. A disseminação de notícias falsas pode causar sérios problemas em áreas cruciais da sociedade, tais como a mídia e a política acarretando consequências graves para a sociedade como um todo. Com isso, surge a importância de uma boa utilização do jornalismo de dados que, em colaboração com outros jornalistas e organizações, podem verificar a veracidade das informações e expor as notícias falsas.

A verificação de *fake news* requer grades equipes com conhecimento específico em diferentes áreas, tornando-se um trabalho manual e exaustivo. Abordagens de Processamento de Linguagem Natural (PLN) e *Deep Learning* têm sido úteis para a detecção de fake news, ainda assim, essa continua sendo uma tarefa desafiadora [Oshikawa et al. 2018]. Além de classificações binárias (*real* ou *fake*), há casos em que as notícias são parcialmente falsas [Oshikawa et al. 2018]. A criação de *datasets* específicos para o problema é um desafio pela necessidade de grandes volumes de dados.

Um desafio adicional das fake news é a definição do que realmente constitui uma fake news. Uma proposta para essa questão é a divisão em três tipos [Rubin et al. 2015]: sensacionalismo, que enfatiza o aspecto emocional da notícia em detrimento da precisão dos fatos; *hoaxing*, que se refere a notícias criadas com o objetivo expresso de enganar o público; e notícias humorísticas, que podem ser confundidas com notícias verdadeiras, mas que têm como objetivo apenas entreter.

Em suma, a crescente informatização e o acesso fácil a informação criaram um ambiente propício para a disseminação de notícias falsas. As *fake news* podem causar danos significativos à sociedade, afetando áreas cruciais como a mídia e a política. Nesse contexto, o jornalismo de dados aparece como um papel importante de verificação e exposição das notícias falsas. Apesar dos avanços na detecção automatizada, este trabalho continua sendo manual e exigente. Portanto, nota-se a importância de parcerias entre jornalistas, empresas e a acadêmia para desenvolver novas técnicas, abordagens e a conscientização dos cidadãos, incentivando-os a verificar e adotar uma postura crítica em relação às informações antes de compartilhá-las.

5. Considerações finais

O jornalismo de dados é uma área importante e em crescimento dentro do jornalismo atual. Com o aumento de informações e dados disponíveis, as empresas de comunicação precisam adotar novas estratégias e ferramentas para lidar com essa quantidade de informações. As ferramentas de *big data* são essenciais para a produção de notícias precisas e completas, assim como são importantes para a coleta e análise de informações. Os jornalistas podem explorá-las para entender melhor as informações, criando matérias mais relevantes, que interessem os leitores. Além disso, o jornalismo de dados pode ajudar a prever tendências e identificar padrões e conexões antes desconhecidos. Ferramentas de IA generativa sem dúvida serão cada vez mais utilizadas na geração de notícias, embora

no estado atual da tecnologia existam grandes desafios a serem superados, como potenciais vieses e alucinações. Podemos dizer que o jornalismo de dados se torna cada vez mais importante para a produção e consumo de notícias em um mundo conectado e informatizado, oferecendo uma vantagem competitiva para as empresas de comunicação que souberem se adaptar a essa nova realidade.

Referências

Aitamurto, T., Sirkkunen, E., and Lehtonen, P. (2011). Trends in data journalism. *Espoo: VTT*, pages 0–27.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., et al. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Bounegru, L. and Gray, J. (2021). *The Data Journalism Handbook: Towards a Critical Data Practice*. Amsterdam University Press.

Broussard, M., Diakopoulos, N., Guzman, A. L., Abebe, R., Dupagne, M., and Chuan, C.-H. (2019). Artificial intelligence and journalism. *Journalism & Mass Communication Quarterly*, 96(3):673–695.

Canavilhas, J. and Jorge, T. d. M. (2022). Fake news explosion in portugal and brazil the pandemic and journalistsrsquo; testimonies on disinformation. *Journalism and Media*, 3(1):52–65.

da Silva, R. N. M., Spritzer, A., and Freitas, C. D. S. (2018). Visualization of roll call data for supporting analyses of political profiles. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 150–157. IEEE.

Fink, K. and Anderson, C. W. (2015). Data journalism in the united states: Beyond the "usual suspects". *Journalism studies*, 16(4):467–481.

Gray, J., Chambers, L., and Bounegru, L. (2012). *The data journalism handbook: How journalists can use data to improve the news*. " O'Reilly Media, Inc.".

Hammond, P. (2017). From computer-assisted to data-driven: Journalism and big data. *Journalism*, 18(4):408–424.

Oshikawa, R., Qian, J., and Wang, W. Y. (2018). A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*.

Pavlik, J. V. (2023). Collaborating with chatgpt: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*, page 10776958221149577.

Rubin, V. L., Chen, Y., and Conroy, N. K. (2015). Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Vicente, P. N. and Flores, A. M. M. (2021). Inteligência artificial e jornalismo. *De que falamos quando dizemos Jornalismo?*, pages 175–194.

Wang, F.-Y., Miao, Q., Li, X., Wang, X., and Lin, Y. (2023). What does chatgpt say: the dao from algorithmic intelligence to linguistic intelligence. *IEEE/CAA Journal of Automatica Sinica*, 10(3):575–579.