Estudo sistemático sobre a literatura de mineração de dados no jornalismo [1]

Eduardo Yuji Yamamoto[2]
Kayla Fernanda de Lima[3]
Universidade Estadual do Centro-Oeste, Guarapuava, PR

RESUMO
O presente texto trata de um estudo sistemático da literatura atualizada sobre a mineração de dados no jornalismo. Para tal, utilizou-se o método de revisão de meta-síntese para sistematizar o conceito e descobrir o seu atual estado de arte. Hoje, o estudo sobre esse conceito é fundamental, já que muitas empresas utilizam essa prática com o intuito de disporem de grandes bancos de dados estruturados para melhorarem a destinação de seus produtos. A questão aqui foi descobrir quais os autores e as publicações sobre o tema são os mais citados (em língua portuguesa e inglesa), quais as relações entre elas, além de um estudo preliminar sobre esse conceito. Esse trabalho faz parte de uma pesquisa em andamento.

PALAVRAS-CHAVE: Inteligência Artificial; Aprendizado de Máquina, Algoritmo; Web Scraping.

Hoje, é cada vez mais frequente o uso de dados de navegação por parte de gigantes do mercado digital para a construção de um perfil de usuário que se torna também foco da produção direcionada de conteúdos jornalísticos.

É nesse contexto que se insere o presente trabalho. A ideia de se pesquisar o conceito de mineração de dados no jornalismo tem a ver com o ambiente de grandes transformações tecnológicas que atingem esse setor. Nos últimos anos, a produção de dados multiplicou vertiginosamente no mundo digital, tornando impossível o trabalho de organização dessas informações sem o uso de filtros (BACHUR, 2021) ou máquinas específicas para essa tarefa.

Gigantes do mercado como *Google, Apple, Facebook, Amazon e Microsoft* dependem diretamente da mineração de dados, a partir dos chamados rastros digitais (*likes*, recensões, histórico de compras em sites), para melhorar o seu desempenho como empresas num mercado tão competitivo. A questão, entretanto, é saber se o jornalismo também tem se servido dessa prática não apenas para integrá-la a um

[1] Trabalho apresentado no Grupo de Trabalho Estudos da Comunicação, evento integrante da programação do 23º Congresso de Ciências da Comunicação na Região Sul, realizado de 13 a 15 de junho de 2024.
[2] Professor do Curso de Comunicação Social do DECS-UNICENTRO, email: yamamoto@unicentro.br
[3] Estudante de Graduação 5º semestre do Curso de Jornalismo do DECS-UNICENTRO, email: kaylaamil2022@gmail.com

negócio que está em vias de automação, mas também para aperfeiçoar a produção de notícias que nascem, precisamente, de dados estruturados.

Como temos observado, o estudo desse conceito tem se dado em interface com campos ainda estranhos ao Jornalismo como a Ciência da Computação, a Engenharia ou a Matemática Aplicada, o que pode indicar uma mutação em sua própria epistemologia. Para Walter Lima Jr. (apud. IOSCOTE, 2021, p. 177):

> [...] há uma necessidade de entendimento, por parte dos jornalistas, sobre o funcionamento dos sistemas tecnológicos que suportam a mídia [...] durante anos, tanto o ensino quanto a prática do Jornalismo estiveram longe dessa perspectiva. O resultado é que a maior parte dos jornalistas crê que utilizar um computador, lidar com e-mails e redes sociais digitais seria o suficiente para entender o jornalismo contemporâneo.

Com o estudo do conceito, buscou-se compreender o interesse dos pesquisadores de Jornalismo e Comunicação sobre o tema: há uma demanda pelo estudo da mineração de dados? Quais são seus autores mais relevantes, quais suas fontes de informação sobre o tema?

Como metodologia, utilizou-se a revisão de literatura por meio da meta-síntese. Ela se baseia em, primeiramente, fornecer informações que possam ser reproduzidas por outros pesquisadores, "apresentando de forma explícita as bases de dados bibliográficos que foram consultadas, as estratégias de busca empregadas em cada base, o processo de seleção dos artigos científicos, os critérios de inclusão e exclusão dos artigos e o processo de análise de cada artigo" (GALVÃO; RICARTE, 2019, p. 58). Em segundo lugar, propor a descrição "sobre um tópico a fim de localizar temas, conceitos ou teorias-chave que forneçam novas ou mais poderosas explicações para o fenômeno sob análise" (IDEM, p. 60).

Assim, realizamos um levantamento de materiais na plataforma *Google Acadêmico* a fim de extrair textos sobre o tema para, posteriormente, fornecer uma síntese dos dados significativos. Para conduzir esse levantamento de materiais de forma eficiente, foram utilizados os seguintes medidores *booleanos*: "mineração de dados", "algoritmos", "aprendizado de máquina" e "base de dados".

Dada a natureza relativamente recente do tema, considerou-se fundamental explorar o amplo espectro de pesquisas disponíveis, tanto em língua inglesa quanto em língua portuguesa. Além disso, atentou-se para a análise daqueles materiais que

apresentassem um maior número de citações buscando, assim, abranger os textos mais citados, considerados os mais influentes no campo acadêmico.

O objetivo desse levantamento foi compreender de maneira abrangente e profunda a prática da mineração de dados no jornalismo, bem como suas implicações e influências na produção e disseminação das notícias. O procedimento teve o seguinte itinerário: após a busca e seleção dos materiais já realizada, foi estabelecido o campo de problematizações a fim de identificar as epistemologias convergentes presentes nos artigos, contribuindo para uma análise mais abrangente e integrada do tema em questão.

Abaixo, apresentamos o levantamento realizado, seguindo a ordem decrescente, por número de citações, em língua portuguesa (Tabela 1), e em língua inglesa (Tabela 2), realizado entre os dias 27 de fevereiro e 12 de março de 2024. Foram utilizados os seguintes marcadores: “journalism”, “data mining” “big data”, “machine learning", “deep learning”, “jornalismo”, “mineração de dados” e “jornalismo de dados”.

Tabela 1 - Levantamento de trabalhos em língua portuguesa

Título	**Local (revista ou livro, dissertação e tese)**	**Autores**	**Ano**	**Número de citações**
Sistematizando conceitos e características sobre o jornalismo digital em base de dados	Livro “Jornalismo digital de terceira geração”	Suzana Barbosa	2007	22
Jornalismo inteligente na era do data mining	Artigo	Walter Teixeira Lima Junior	2016	20
Mineração de dados e textos e suas possibilidades aplicadas ao processo de produção da notícia	Artigo	Pablo Barbosa Walter Teixeira Lima Junior	2007	1

Fonte: Dos próprios autores.

Tabela 2: Levantamento de trabalhos em língua inglês

Título	**Nome da revista**	**Autores**	**Ano**	**Número de citações**
Computational journalism	Artigo	Sarah Cohen, James T. Hamilton, Fred Turner	2011	379
Algorithmic Journalism—Current Applications and Future Perspectives	Artigo	Efthimis Kotenidis Andreas Veglis	2021	42
Mining social media for newsgathering: A review	Artigo	Arkaitz Zubiaga	2019	42

Fonte: Dos próprios autores.

Inicialmente, é imprescindível destacar que o processo de mapeamento dos materiais conduzido para esta pesquisa revelou uma ampla gama de textos pertinentes ao tema em questão. Contudo, devido ao pouco espaço deste formato de apresentação, foi necessário restringir o escopo de análise aos seis materiais mais relevantes, ordenados por número de citação. Os materiais selecionados se complementam, construindo uma abordagem completa dentro do conceito de mineração de dados no jornalismo, dialogando entre si, além de construir uma linha do tempo sobre a temática, a partir da data de publicação de cada texto, o que enriqueceu a perspectiva histórica e evolutiva do tema.

No âmbito dos materiais mais citados em língua portuguesa, observou-se uma lacuna temporal que não acompanhou o desenvolvimento tecnológico dos anos recentes, sendo o artigo mais atual de 2016. Por outro lado, quando analisamos os materiais em língua inglesa, outra realidade se apresenta: o texto mais recente remonta a 2021.

Mesmo com a lacuna temporal entre as pesquisas, seus tópicos se assemelham consistindo em quatro temáticas principais: a conceituação da mineração de dados e base de dados, a defesa da tecnologia como prática social e elemento intrínseco ao jornalismo - mesmo em materiais anteriores à década atual, esse conceito já era utilizado por BARBOSA (2007) -, a defesa da mineração de dados como ferramenta de pesquisa e verificação de fatos, além de questionamentos sobre a utilização de máquinas computacionais para a personalização das notícias, de forma a garantir maior visibilidade do público de acordo com suas preferências armazenadas nos grandes bancos de dados.

Para além disso, é perceptível a intensificação em discussões de teor ético em relação ao tema com o decorrer dos anos de pesquisa, visto o desenvolvimento das inteligências artificiais.

> Além das óbvias limitações técnicas, muitas considerações editoriais e éticas surgiram, sinalizando que o cenário da automação é simultaneamente muito promissor e muito desafiador também. À medida que o ceticismo em relação às preocupações com privacidade e trabalho jornalístico atinge o auge, é importante que os algoritmos permaneçam o mais transparentes e bem regulamentados possível, para continuar seu desenvolvimento em harmonia com os valores jornalísticos tradicionais e, em última análise, realizar seu potencial ajudando os jornalistas a superar algumas das limitações fundamentais da profissão e avançar o trabalho jornalístico além do que é atualmente possível. (KOTENIDIS, 2021)

No contexto dos autores mais frequentemente citados, uma diversidade de pesquisadores emerge tanto em língua portuguesa quanto inglesa. Algumas figuras se destacam pela frequência de suas citações, denotando uma contribuição substancial para o campo. Um exemplo é Walter Teixeira Lima Junior, que figura em dois dos artigos mais citados em língua portuguesa. Apesar de "Computational journalism" ser o texto mais citado em língua inglesa, com mais de trezentas e setenta e nove citações, os materiais mais referenciados em língua portuguesa prescindem dessa obra. Dentro do universo bibliográfico utilizado por pesquisadores brasileiros para fundamentar suas investigações no âmbito da mineração de dados, destaca-se a obra "Data mining: concepts and techniques", de Jiawei Han e Micheline Kamber, publicada em 2001. De uma maneira geral, o tema reúne pesquisas tanto do campo da comunicação, como da ciência da computação, se mantendo entrelaçado às ciências sociais e ciências naturais voltada ao fenômeno da cognição.

De maneira ampla, observa-se uma baixa demanda por parte dos pesquisadores brasileiros no contexto das pesquisas recentes sobre o tema, com uma incidência muito maior de trabalhos em língua inglesa, tanto em termos de data de publicação quanto de temáticas abordadas – o que reflete um pouco a grande diferença econômica e tecnológica entre a realidade brasileira e dos países anglófonos.

REFERÊNCIAS

BACHUR, João Paulo. Desinformação Política, Mídias Digitais e Democracia: como e por que as fake news funcionam? RDP, Brasília, v. 18, n. 99, p. 436-469, 2021.

BARBOSA, Pablo; LIMA JÚNIOR, Walter Teixeira. Mineração de dados e textos e suas possibilidades aplicadas ao processo de produção da noticia. ANAIS DO 5º ENCONTRO NACIONAL DE PESQUISADORES EM JORNALISMO, Universidade Federal de Sergipe, 2007.

FARIAS, Marcello Tenorio; ANGELUCI, Alan César Belo Angeluci; PASSARELLI, Brasilina. Web Scraping e ciência dos dados na pesquisa aplicada em Comunicação: um estudo sobre avaliações online. Revista Observatório, Palmas, v. 7, n. 3, p. 1-22, 2021.

IOSCOTE, Fabia Cristiane. Jornalismo e Inteligência Artificial: tendências nas Pesquisas Brasileiras entre 2010 e 2020. Novos Olhares, v. 10, n.2, p. 162-182, 2021.

GALVÃO, Maria Cristiane Barbosa; RICARTE, Ivan Luiz Marques. Revisão sistemática da literatura: conceituação, produção e publicação. Logeion - Filosofia da informação, Rio de Janeiro, v. 6, n. 1, p. 57-73, 2019.

LIMA JÚNIOR, Walter Teixeira. Jornalismo inteligente na era do data mining. Revista do Programa de Pós-Graduação da Faculdade Cásper Líbero, n.8, p. 119-126, 2006.

BARBOSA, Suzana. Sistematizando conceitos e características sobre o jornalismo digital em base de dados. Jornalismo digital de terceira geração.

COHEN, Sarah; HAMILTON, James T.; TURNER Fred. Computational journalism. ACM Portal. v. 54, n. 10, p. 66-71, 2011.

KOTENIDIS, Efthimis; VEGLIS, Andreas. Algorithmic Journalism — Current Applications and Future Perspectives. Journalism and Media. n.2, p. 244-257, 2021.

ZUBIAGA, Arkaitz. Mining social media for newsgathering: A review. Elsevier. v.13, 2019.