QUANTIFICAÇÃO DAS CITAÇÕES DAS INSTITUIÇÕES PÚBLICAS DE ENSINO NOS DEBATES SOBRE INFLAÇÃO NA GRANDE MÍDIA

Gabriel Messias Corradi de Souza (1) – Luiz Guilherme Hiel Drumond Silveira

João Vitor de Oliveira Diniz – Luiz Guilherme Hiel Drumond Silveira

RESUMO

O presente trabalho tem como objetivo quantificar as citações de instituições públicas de ensino em reportagens sobre inflação publicadas na grande mídia a partir de 2014. Para isso, desenvolveu-se um robô *scraper* em *Python*, com *Scrapy* e *spaCy*, que coleta e processa textos extraídos dos principais veículos de mídia do Brasil na Internet. No desenvolvimento do trabalho, a identificação dos autores das reportagens entre diferentes veículos demandou estratégias híbridas com auxílio do *spaCy*. Os resultados mostram também que há concentração significativa de publicações em poucos veículos, especialmente UOL e Folha de São Paulo, e que o robô é válido para cumprir a tarefa.

Palavras-chave: Inflação. *Scraper*. Jornalismo.

1 INTRODUÇÃO

O Jornalismo é uma das instituições sociais de maior importância em uma sociedade liberal. Em um país com liberdade de imprensa, a sua função serve como mediadora entre o Estado e a sociedade (FOX, 2013). A ciência, com seus métodos, permite interpretar a realidade de forma objetiva. Sendo assim, é de fundamental importância que os jornalistas, para o bem da qualidade da informação que será repassada, entrevistem e se baseiem na visão de cientistas. É o que aconteceu, por exemplo, durante o período da COVID-19, os principais veículos de mídia entrevistaram e citaram pesquisadores brasileiros na construção de suas reportagens (REVADAM, 2023).

(1) Aluno Bolsista; Curso: Bacharelado em Engenharia de Controle e Automação; Instituição: Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais - Campus Sabará; Modalidade da bolsa: AT-NS; Fonte Financiadora: Programa Institucional de Fomento a Bolsas de Pesquisa do IFMG Campus Sabará.

Estudantes: Bacharelado em Engenharia de Controle e Automação, Campus Sabará, IFMG – Coordenador do projeto: Mestre em Engenharia Elétrica l, Campus Sabará, IFMG

A produção científica do Brasil é realizada por instituições públicas, mais especificamente, por instituições públicas de ensino (IPEs) (ALMEIDA, 2021). Considerando essa premissa, espera-se que os principais pesquisadores brasileiros, presentes nas IPEs, pelo bem do processo de informar a população, sejam citados ou entrevistados em reportagens de todos os segmentos jornalísticos, como a Economia, por exemplo. Pois um jornalismo econômico que se baseia apenas em opinião, com viés e sem pluralidade leva a opinião pública a ter uma percepção distorcida da realidade econômica (BOYDSTUN; HIGHTON; LINN, 2018).

Como em artigos jornalísticos a instituição em que o entrevistado ou o citado é sempre citada, este trabalho propõe: (i) a criação de um robô *scraper* para realizar uma busca no Google com o a palavra-chave "inflação"; e (ii) utilizar processamento de linguagem natural para extração do nome do autor da reportagem, o entrevistado, o citado, e as suas instituições. A proposta tem como objetivo quantificar e categorizar o número de citações das instituições de ensino públicas brasileiras no debate da grande mídia brasileira.

2 DESENVOLVIMENTO

Nesta seção serão apresentados os trabalhos relacionados juntamente com a fundamentação teórica, bem como os materiais e métodos utilizados neste trabalho.

2.1 Trabalhos Relacionados e Fundamentação Teórica

A forma como a grande mídia aborda temas econômicos influencia diretamente a confiança dos agentes econômicos e, consequentemente, suas decisões, podendo gerar efeitos como crises especulativas e fuga de capitais (PASSOS; GONÇALVES, 2024). O jornalismo exerce um papel de autoridade informacional, moldando percepções sociais sobre a conjuntura econômica (TIWARI, 2021). Nos Estados Unidos, por exemplo, a cobertura tende a refletir os interesses das classes mais ricas, com variações de tom conforme seus rendimentos, resultado não de viés intencional, mas de abordagens metodológicas reducionistas baseadas em indicadores isolados, como PIB e taxa de desemprego (JACOBS et al., 2021).

No contexto brasileiro, Souza e Neto (2019) analisaram a cobertura da reforma da previdência em diferentes períodos, constatando baixa representatividade de vozes acadêmicas nos principais veículos de comunicação, o que contrasta com a predominância da produção científica concentrada nas universidades públicas. Em contrapartida, Revadam (2023) observou expressivo crescimento do jornalismo científico durante a pandemia de COVID-19, com ampla veiculação de matérias de caráter técnico e científico em telejornais nacionais.

Paralelamente, o avanço tecnológico permitiu o uso de robôs *scrapers* integrados ao Processamento de Linguagem Natural (PLN), ferramenta da Inteligência Artificial voltada à compreensão e análise da linguagem humana (O'REILLY, 2020). Em *Python*, destacam-se o *Scrapy*, voltado à coleta estruturada de dados *web* (CAMPOS, 2021), e o *spaCy*, especializado em análise linguística eficiente (VASILIEV, 2020). Experimentos como o de Fernández-Cruz e Moreno-Ortiz (2020) demonstram a eficácia dessa integração, utilizando buscadores e *scripts* automatizados para coletar e processar dados textuais. De forma complementar, Chaitanya e Shetty (2023) empregaram o *Selenium* para possibilitar a interação de *scrapers* com páginas dinâmicas e sistemas baseados em geolocalização, ampliando a autonomia e a aplicabilidade desses agentes.

2.2 Materiais e Métodos

O desenvolvimento foi realizado em um ambiente com sistema *Windows* 10 (64 bits), processador Intel ~2,3 GHz e 8 (oito) GB de RAM, utilizando o *Visual Studio Code* com *Python* 3.11 e as bibliotecas *Scrapy*, *spaCy*, Pandas, JSON e CSV. O robô foi implementado em *Python* com o *framework Scrapy*, responsável pela coleta estruturada de dados, e o *spaCy*, empregado para o processamento linguístico. A integração da *API (Application Programming Interface) Google Custom Search* permitiu a busca automatizada do termo "inflação" e o armazenamento dos resultados — títulos e HTML — em arquivos JSON. Em seguida, o *spaCy* realizou a extração de entidades nomeadas, como autores, pessoas, instituições e datas. Por fim, o sistema foi validado por meio da verificação da quantidade de resultados retornados por veículo de mídia e da consistência das extrações realizadas.

2.3 Resultados Parciais

A etapa final consistiu na validação do robô, iniciada pela verificação da capacidade de realizar buscas no Google. Devido à limitação da *API*, que retorna até 10 resultados por consulta, foi realizada a comparação entre o HTML obtido e o original, confirmando a precisão das extrações para todos os veículos da plataforma *Media Ownership Brazil* (2017). Embora a exibição fosse restrita, a *API* permitiu identificar o volume total de publicações, revelando forte concentração nos portais UOL e Folha de São Paulo, com mais de 8 (oito) e 7 (sete) milhões de resultados, respectivamente, entre 2014 e 2025.

A extração de autoria apresentou maior complexidade por conta da ausência de padronização entre sites. Para superar essa limitação, foram aplicadas técnicas automatizadas de detecção de padrões — como análise de prefixos textuais, proximidade de datas e estruturas específicas do HTML —, garantindo a robustez do processo de identificação dos autores.

3 CONSIDERAÇÕES FINAIS

O jornalismo desempenha um papel essencial na sociedade, uma vez que exerce funções fundamentais para o fortalecimento da democracia, a formação da opinião pública e a difusão de informações. Nesse contexto e a partir dos resultados apresentados, conclui-se que o robô está validado para quantificar a presença das IPEs nas reportagens da grande mídia. Como próximos passos, pretende-se não apenas quantificar as IPEs, mas também construir o cenário quantificando de quais são as instituições, citados, entrevistados e autores das reportagens por todo período. Como trabalhos futuros pretende-se empregar técnicas de análise de sentimentos em cada reportagem levantada.

REFERÊNCIAS

ALMEIDA, S. **A ciência, as universidades e o futuro do país.** Disponível em: <https://ufmg.br/comunicacao/noticias/a-ciencia-as-universidades-e-o-futuro-do-pais>. Acesso em: 9 ago. 2024.

BOYDSTUN, A.; HIGHTON, B.; LINN, S. Assessing the Relationship between Economic News Coverage and Mass Economic Attitudes. **Political Research Quarterly**, v. 71, n. 4, p. 989–1000, 2018.

CAMPOS, Sandro Luís Brandão. *Desenvolvimento de ferramenta para coleta e análise de dados web utilizando Scrapy e Python.* 2021. Trabalho de Conclusão de Curso (Mestrado Profissional em Propriedade Intelectual e Transferência de Tecnologia para Inovação) – Universidade Federal de Mato Grosso, Cuiabá, 2021. Disponível em: https://profnit.org.br/wp-content/uploads/2021/04/UFMT-Sandro-Luis-Brandao-Campos-TCC.pdf. Acesso em: 4 set. 2025.

CHAITANYA, A.; SHETTY, J. Food Image Classification and Data Extraction Using Convolutional Neural Network and Web Crawlers. *Procedia Computer Science*, v. 281, p. 143-152, 2023.

FERNÁNDEZ-CRUZ, J.; MORENO-ORTIZ, A. Building the Great Recession News Corpus (GRNC): A contemporary diachronic corpus of economy news in English. **Research in Corpus Linguistics,** n. 8, p. 28-45, 2020.

FOX, C. Public reason, objectivity, and journalism in liberal democratic societies. *Res Publica*, v. 19, p. 257-273, 2013. DOI: https://doi.org/10.1007/s11158-013-9226-6.

JACOBS, A.; MATTHEWS, J.; HICKS, T.; MERKLEY, E. Whose News? Class-Biased Economic Reporting in the United States. **American Political Science Review**, v. 115, n. 3, p. 1016-1033, 2021.

O'REILLY. *Machine Learning & Data Science Blueprints for Finance: From Building Trading Strategies to Robo-Advisors Using Python.* O'Reilly Media, 2020. 1. ed. p. 347.

PASSOS, G.; GONÇALVES, R. O efeito da cobertura da mídia no gerenciamento de resultados: evidências das companhias brasileiras. **Revista Contabilidade e Finança**, v. 35, n. 1869, 2024.

REVADAM, R. **Coronavírus: quando a ciência ganha as manchetes uma análise das reportagens científicas em tempo de pandemia.** Campinas: UNICAMP, 2023. p. 121.

SOUZA, T.; NETO, A. O Jornalismo Econômico e as Vozes Que Falaram nos Jornais nos Anos de Debate das Reformas da Previdência. **Revista Observatório**, v. 5, n. 6, p. 634-667, 2019.

TIWARI, P. Effect of Media on the Behaviour of Investors and Stocks. **Turkish Online Journal of Qualitative Inquiry**, v. 12, n. 6, p. 1667-11673, 2021.

VASILIEV, Yuli. *Natural Language Processing with Python and spaCy: A Practical Introduction.* San Francisco: No Starch Press, p. 2, 2020. ISBN 9781718500525.