Comparativo de Modelagem de Tópicos: um estudo de caso de relatos de publicações predatórias

Fernando A. do Carmo[1], **Marcelino S. da Silva**[1],
Antonio F. L. Jacob Junior[2], **Fábio M. F. Lobato**[1]

[1]Instituto de Engenharia e Geociências
Universidade Federal do Oeste do Pará – Santarém – PA – Brazil

[2]Departamento de Engenharia da Computação
Universidade Estadual do Maranhão – São Luís – MA – Brazil

antoniojunior@professor.uema.br, fabio.lobato@ufopa.edu.br

Resumo. *A necessidade imperativa de publicação deu origem a uma indústria que oferece um atalho preocupante para pesquisadores sob pressão por novas publicações, prometendo acesso aberto e respostas extremamente rápidas, esses veículos foram alcunhados de predatórios por Jeffrey Beall em 2010. Essas vias de publicação se tornam alternativas atraentes para cumprir as exigências mínimas na pós-graduação. O presente trabalho comparou diferentes algoritmos de Modelagem de Tópicos para avaliar a percepção do conhecimento deste fenômeno por pesquisadores brasileiros. Os resultados são cruciais para identificar como os pesquisadores selecionam periódicos para publicações e apontar atividades que podem auxiliar no combate dessas práticas.*

1. Introdução

Publique ou pereça, este mote está arraigado aos pesquisadores mundo afora [Tian et al. 2016]. Em especial, o pesquisador brasileiro não apenas tem a necessidade de publicar como também precisa selecionar periódicos que estejam listados no Sistema Qualis, da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) [Perlin et al. 2018]. A imperatividade da publicação veio acompanhada do surgimento de uma indústria que oferece um atalho fatídico para isso, prometendo acesso aberto *(open-access)* e respostas extremamente rápidas, estes periódicos foram alcunhados de *"predatórios"* pelo bibliotecário Jeffrey Beall em 2010 [Krawczyk and Kulczycki 2021].

Estes periódicos são caracterizados pelo envio de *e-mails* assediosos, ausência de revisão por pares, inserção de nomes no corpo editorial sem autorização, cobrança de taxas de revisão e publicação e disponibilização de alegações falsas sobre indexação e fator de impacto [Wang and Soler 2021]. Não obstante os alertas para editoras predatórias, esses atalhos de publicação se tornam alternativas atraentes para cumprir as exigências mínimas, principalmente quando o periódico-alvo está indexado em sistemas como o Qualis/CAPES, que tornam essa publicação bem vista por contar para a avaliação do Programa de Pós-Graduação (PPG). Agências de fomento e instituições de pesquisa utilizam tais aspectos para avaliar a qualidade da produção científica dos pesquisadores. Além disso, a classificação dos periódicos no sistema é levada em consideração na hora de avaliar a produção científica dos pesquisadores e, consequentemente, influencia na concessão de bolsas, financiamentos e outros tipos de apoio à pesquisa [Santos and Rabelo 2017].

Considerando a relevância desta problemática, o Grupo de Estudo e Pesquisa em Computação Aplicada da Universidade Federal do Oeste do Pará conduziu um *survey* com 3.067 pesquisadores de diferentes PPGs do país. Dentre as variáveis coletadas, foram incluídas perguntas abertas sobre como pesquisadores(as) percebiam as publicações predatórias. Visando extrair conhecimento de tais relatos e, baseando-se no trabalho de [Baumer et al. 2017], comparamos métodos de modelagem de tópicos e teoria fundamentada. Para tal, empregamos os seguintes métodos de modelagem de tópicos: *Latent Dirichlet Allocation* (LDA), *Latent Semantic Analysis* (LSA) e *Non-Negative Matrix Factorization* (NMF). Deste modo, foi possível estimar qual algoritmo se destacou como o mais adequado para extrair conhecimento dos dados textuais coletados.

2. Procedimentos Metodológicos

Foi adotado um modelo de pesquisa quantitativa-qualitativa *(Survey)*. O presente estudo se baseou na análise de dados de um questionário contendo 43 itens que incluíam dados sociais, geográficos, atuação e experiências acadêmicas de profissionais e estudantes atuantes no cenário de pesquisa e desenvolvimento no Brasil, majoritariamente encontrados nos PPGs das universidades. O presente estudo concentrou-se na análise dos questionamentos abertos utilizando técnicas de modelagem de tópicos e Teoria Fundamentada. Especificamente, foram analisados relatos de experiência com predatismo (317 respostas); e relatos de como o(a) pesquisador(a) escolhiam os periódicos-alvo para publicação (2.132 respostas). A preparação dos dados consistiu na aplicação dos seguintes passos de pré-processamento dos textos: remocão de *stopwords*, remoção de pontuação, remoção de caracteres especiais e transformação para caixa baixa. Diferentes algoritmos de modelagem de tópicos foram comparados, a saber: *Latent Dirichlet Allocation*, *Latent Semantic Analysis* e *Non-Negative Matrix Factorization*. Os métodos foram implementados utilizando a biblioteca *scikit-learn*, com uma parametrização de 5 tópicos por variável e 10 termos por tópico.

3. Resultados e Discussão

Dos 3.067 perfis de pesquisadores coletados inicialmente, foram analisados 2.449 relatos, uma vez que alguns participantes optaram por não responder a essas perguntas específicas da pesquisa. As análises mostram que o algoritmo de modelagem não supervisionado *Non-Negative Matrix Factorization*, testado com diferentes configurações (número de tópicos e número de termos por tópico), alcançou resultados mais satisfatórios. Este método se mostrou mais eficiente para este estudo por conseguir produzir tópicos mais coerentes em detrimento aos outros métodos utilizados, além de possuir um melhor desempenho em termos de mineração de tópicos de texto curto, assim como destacado por [Costa et al. 2022]. Essa constatação foi avaliada qualitativamente por um grupo de analistas independentes. O processo de análise consistiu em calibrar os algoritmos e avaliar qual deles se destacava na identificação do maior número de tópicos. Por questões de tamanho do manuscrito, a seguir são apresentados os resultados para o NMF.

As Tabelas 1 e 2 apresentam os tópicos mais aparentes nestas duas variáveis da pesquisa que foram definidas na Seção anterior. Os tópicos apresentados na Tabela 1 apresentam os principais termos descritos pelos participantes da pesquisa que estão relacionados com suas experiências com publicações predatórias. Dentre os principais tópicos

encontrados é notável uma grande tendência dos pesquisadores mencionarem termos relacionados ao assédio por parte dos periódicos, que é caracterizado, muitas vezes, por meio de convites encaminhados aos pesquisadores com frequência e com promessas tendenciosas, tais como de publicação rápida, suposta revisão por pares e falsos índices de fator de impacto. Além disso, é perceptível termos relacionados a falta de conhecimento do pesquisador com o tema.

Tabela 1. Tópicos sobre experiência com predatismo.

Tópico	Termos
Convites	artigo publicado periodico predatorio recebi mail taxa aceito journal convite
Desconhecimento	revista predatoria qualis trabalho nome publicou apos lista sabia tratava
Evitar	periodicos publicacoes revistas sei predatorias artigos trabalhos nenhuma publicar ainda
Processo	publicacao rapida qualis revisao avaliacao taxa solicitacao processo oportunidade promessa

Tabela 2. Tópicos sobre passos escolha periódico.

Tópico	Termos
Seleção Inicial	listo area periodicos consulto filtro lista conhecidos procuro atuacao acordo
Critérios de Qualidade	melhor seleciono base qualis recomendados filtro menor indicacao tematica basicamente
Consultas com Pares	discuto colegas lista recomendados filtro orientador periodicos melhor orientadores professores
Eficiência e Custo	tempo resposta menor publicacao custo verifico considero periodicos

De modo geral, os dados sobre as experiências com predatismo sugerem um certo desconhecimento dos pesquisadores brasileiros sobre a prática assediosa dos periódicos. Esta falta de atenção ao conceito deve-se a conjuntura de fatores dos quais eles estão inseridos, sendo eles: a necessidade de volume de publicação principalmente pelos pesquisadores de nível maior que graduação ou pelos que estão no processo de obtenção de títulos; o desejo de um pesquisador inexperiente em ver o seu estudo publicado; a falta de comunicação e apoio por parte das instituições e pesquisa com ampla divulgação e orientação acerca do tema. Os periódicos predatórios aproveitam-se desta conjuntura, explorando precisamente essas vulnerabilidades. Destacar esses atores negativos é um dos passos primordiais no combate a essa atividade.

A Tabela 2 destaca os tópicos mais aparentes acerca dos passos que o pesquisador costuma seguir ao publicar um novo trabalhos. Os tópicos sugerem que os pesquisadores costumam consultar listas de periódicos predatórios e buscam por critérios de qualidade dos periódicos, como fator de impacto e *H-Index*. Outra característica encontrada nas análises está relacionada a discussões sobre onde publicar, onde o(a) pesquisador(a) costuma consultar seus pares e/ou seu orientador(a), além de procurar apoio com o próprio programa de pós-graduação ao qual possui vínculo. Em suma, as análises mostram, que os pesquisadores que seguem os passos destacados pelos termos na Tabela 2 para suas publicações já tiveram alguma experiência com publicações predatórias. As discussões acerca de como evitar esses veículos são cruciais para ajudar o pesquisador na tomada de decisão e no combate dessa prática no Brasil, uma vez que tais práticas geram consequências que refletem diretamente na carreira e credibilidade do autor.

4. Considerações Finais

O presente trabalho apresentou uma análise comparativa entre diferentes algoritmos de modelagem de tópicos para extrair conhecimento de dados textuais, foram analisados relatos de experiências com predatismo por pesquisadores brasileiros e os principais passos que estes seguem para uma nova publicação. Os resultados mostram que o algoritmo NMF obteve o melhor desempenho em relação aos outros métodos comparados. Este resultado foi validado por meio da teoria fundamentada, onde os autores leram os relatos e compararam com os tópicos encontrados pelos algoritmos comparados.

A partir dos resultados obtidos neste estudo foi possível identificar quais as práticas estão sendo utilizadas para compartilhar informes sobre predatismo, identificar a vulnerabilidade em que os pesquisadores se encontram e ainda, apontar atividades que podem auxiliar no combate dessas práticas. As estratégias adotadas para obtenção de informações necessárias para o desenvolvimento do estudo, resultaram em dados satisfatórios. Mesmo tendo conhecimento do tamanho da comunidade científica do país, as respostas obtidas nos permitem ter um norte das questões que estão envolvidas no cenário geral. Como trabalhos futuros, pretendemos incluir *word embeddings* como BERTopic e também analisar o coeficiente de coerência como medida de avaliação quantitativa.

Agradecimentos

Este trabalho foi apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)-DT-308334/2020; epela Fundação Amazônia de Amparo a Estudos e Pesquisas (FAPESPA) PRONEM-FAPESPA/CNPq nº 045/2021.

Referências

Baumer, E. P., Mimno, D., Guha, S., Quan, E., and Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 68(6):1397–1410.

Costa, G., Couto, D., Junior, A. J., and Lobato, F. (2022). Feminismo e redes sociais online: uma análise de tweets sobre o dia internacional da mulher. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 169–180, Porto Alegre, RS, Brasil. SBC.

Krawczyk, F. and Kulczycki, E. (2021). How is open access accused of being predatory? the impact of beall's lists of predatory journals on academic publishing. *The Journal of Academic Librarianship*, 47(2):102271.

Perlin, M. S., Imasato, T., and Borenstein, D. (2018). Is predatory publishing a real threat? evidence from a large database study. *Scientometrics*, 116:255–273.

Santos, L. R. and Rabelo, D. M. R. d. S. (2017). ProduÇÃo cientÍfica: AvaliaÇÃo, ferramentas e indicadores de qualidade. *PontodeAcesso*, 11(2):3–33.

Tian, M., Su, Y., and Ru, X. (2016). Perish or publish in china: Pressures on young chinese scholars to publish in internationally indexed journals. *Publications*, 4(2).

Wang, Y. and Soler, J. (2021). Investigating predatory publishing in political science: a corpus linguistics approach. *Applied Corpus Linguistics*, 1(1):100001.