MODELAGEM AUTOMÁTICA DE TÓPICOS EM TEXTOS JORNALÍSTICOS REFERENTES A POLUIÇÃO MARINHA

Gustavo Presa Rosa[1], Rodrigo Machado[2], Merisandra Côrtes de Mattos[3]

Resumo: Diante do problema mundial de poluição marinha, o monitoramento acerca desta situação torna-se indispensável. Uma fonte de dados com potencial de fornecer conhecimentos relevantes em relação à sociedade, são as mídias digitais. Portanto, esta pesquisa tem como objetivo analisar textos jornalísticos nacionais publicados em espaços digitais relacionados a poluição marinha por meio da modelagem automática de tópicos utilizando os algoritmos *Latent Dirichlet Allocation* (LDA), *Hierarchical Dirichlet Process* (HDP) e *Structural Topic Model* (STM) e comparar os resultados de coerência semântica obtidos por cada modelo. Os resultados atingidos mostram variações na geração de tópicos semanticamente coerentes de acordo com o valor de tópicos gerado pelo modelo. Por fim, a pesquisa evidencia uma prevalência de notícias relacionadas a poluição por lixo plástico no Brasil.

Palavras-chave: Modelagem de tópicos. Ciência de Dados. Poluição Marinha

ABSTRACT: Faced with the worldwide problem of marine pollution, monitoring this situation becomes indispensable. A source of data with the potential to provide relevant knowledge in relation to society is digital media. Therefore, this research aims to analyze national journalistic texts published in digital spaces related to marine pollution through automatic modeling of topics using the algorithms Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP) and Structural Topic Model (STM) and compare the semantic coherence results obtained by each model. The achieved results show variations in the generation of semantically coherent topics according to the value of topics generated by the model. Finally, the research shows a prevalence of news related to plastic waste pollution in Brazil.

[1]Curso de Ciência da Computação, Grupo de Pesquisa em Inteligência Artificial Aplicada, Universidade do Extremo Sul Catarinense (Unesc), Criciúma - Santa Catarina - Brasil. gustavopr89@gmail.com.
[2]Coorientador, Curso de Ciências Biológicas, Universidade do Extremo Sul Catarinense (Unesc), Criciúma - Santa Catarina - Brasil. rodrigomachado@unesc.net.
[3]Orientadora, Curso de Ciência da Computação, Grupo de Pesquisa em Inteligência Artificial Aplicada, Universidade do Extremo Sul Catarinense (Unesc), Criciúma - Santa Catarina - Brasil. mem@unesc.net.

Keywords: Topic modeling. Data Science. Marine pollution.

1 INTRODUÇÃO

O mundo produz mais de 380 milhões de toneladas de plástico todos os anos, que podem acabar como poluentes, entrando no ambiente natural e nos oceanos (RICHIE, 2019). O aumento da produção em larga escala do plástico e o seu uso na sociedade levaram a um acúmulo massivo de resíduos nos oceanos, mares e rios. O acúmulo de destas partículas plásticas no ambiente marinho está poluindo cada vez mais o planeta, isso pois o plástico, em sua maioria, não é biodegradável, sendo assim, ele não se decompõe naturalmente de uma forma que não seja prejudicial ao meio ambiente. Considerando-se isso, são necessárias medidas relevantes para que seja possível enfrentar a crise global da poluição marinha (UNEP, 2021). Entretanto, mesmo com a poluição marinha a cada dia se tornando um problema mais evidente, a conduta da população em geral não se mostra afetada (LEE et al., 2022, tradução nossa).

Uma das estratégias para monitoramento é compreender as informações referente a poluição marinha disponibilizadas publicamente na mídia, que desempenha um papel crucial na representação e influência da opinião pública nas ações sociais (FORLEO; ROMAGNOLI, 2021, tradução nossa). Uma das maneiras de compreender o que está sendo veiculado nas mídias digitais se dá por meio do Processamento de Linguagem Natural (PLN).

O PLN consiste no desenvolvimento de ferramentas para geração, compreensão e análise automática da linguagem natural humana, disponibilizadas na forma de textos e sons que possam ser interpretados computacionalmente por meio de algoritmos de aprendizado de máquina (GONZALEZ; LIMA, 2003; ZEROUAL; LAKHOUAJA, 2018). Dentre as técnicas de PLN tem-se a modelagem de tópicos, que organiza e resume conteúdos de grandes volumes de dados e informações, denominados de *corpus* de dados, por meio de algoritmos não supervisionados de aprendizado de máquina (BLEI, 2012). Os pesquisadores na área, considerando que

os seres humanos não possuem a capacidade de leitura do quantitativo de textos publicados digitalmente em um determinado domínio do conhecimento, desenvolveram algoritmos que analisam as palavras nos textos e como elas se relacionam. Assim, esses algoritmos identificam tópicos em um conjunto de documentos, o que possibilita a organização, gerenciamento e resumo de textos digitais (BLEI, 2012; ZHU et al., 2012).

Entre os diversos algoritmos aplicados para a modelagem de tópicos tem-se o Alocação Latente de Dirichlet, do inglês *Latent Dirichlet Allocation* (LDA), o Processo Hierárquico de Dirichlet, do inglês *Hierarchical Dirichlet Process* (HDP), e o Modelo de Tópicos Estrutural, do inglês *Structural Topic Model* (STM). Estes são alguns dos algoritmos existentes, os quais são aplicados nesta pesquisa e que segundo Teh et al. (2022) possuem ampla utilização.

O LDA tem como propósito encontrar tópicos-chave inseridos em coleções de registros (BLEI, 2012). Estes tópicos-chave são identificados a partir das relações semânticas encontradas em um conjunto de dados pré-processados (OTERO; GAGO; QUINTAS, 2021, tradução nossa). O LDA é um dos modelos probabilísticos generativos mais utilizados, que emprega uma abordagem bayesiana e considera que os documentos presentes em um *corpus* se referem a misturas aleatórias de tópicos latentes. Posteriormente, cada tópico passa a ser caracterizado por uma distribuição de palavras que compreendem cada um dos documentos (BLEI, 2012).

O HDP é um modelo bayesiano não parametrizado utilizado quando se deseja criar um agrupamento de tópicos a partir de múltiplos grupos de dados. Este algoritmo compartilha com o LDA a capacidade de permitir números incontáveis de empates multinominais, podendo gerar uma quantidade infinita de tópicos. Diferencia-se do LDA pela sua capacidade de determinar dinamicamente o número ideal de tópicos a ser definido durante a execução do modelo.

O STM é um algoritmo bayesiano de tópicos generativos derivado do LDA (ROBERTS et al., 2013), que fornece diversos recursos como a exploração de tópicos, a avaliação de incertezas e a visualização da quantidade de interesse em um

determinado assunto. O STM possui como diferenciais a possibilidade de correlação entre os tópicos e a distribuição prévia dos tópicos estabelecida pela covariável de interesse. As covariáveis adicionais que o modelo suporta provêm da possibilidade de adicionar informações relevantes durante o procedimento de inferência de dados, alterando a estrutura na criação de tópicos.

Cada um desses modelos foram escolhidos para serem aplicados a pesquisa, pois possuem características e abordagens próprias para a identificação de tópicos em um conjunto de documentos. Enquanto o LDA é amplamente utilizado e possui uma base sólida na literatura (BLEI, 2012), o HDP oferece a vantagem de inferir automaticamente o número de tópicos, adaptando-se melhor a conjuntos de dados complexos Teh et al. (2022). Já o STM permite a inclusão de metadados e a modelagem da estrutura de dependência entre palavras (ROBERTS et al., 2013). Ao comparar os resultados desses modelos, é possível obter uma visão mais abrangente e confiável dos tópicos identificados.

Diante das diferentes abordagens técnicas apresentadas pelos modelos, a presente pesquisa tem como objetivo analisar por meio da métrica de coerência semântica a modelagem de tópicos gerada pelos algoritmos LDA, HDP e STM em textos jornalísticos referentes a poluição marinha. Busca-se também, por meio da modelagem de tópicos identificar e analisar quais são os principais tópicos relacionados a poluição marinha discutidos nos textos jornalísticos analisados.

2 TRABALHOS CORRELATOS

Na literatura, existem trabalhos nacionais e internacionais relacionados a modelagem de tópicos que permeiam o desenvolvimento desta pesquisa. Dentre eles, pode-se citar o estudo de Keller e Wyles (2021) que aplicou o algoritmo de modelagem de tópicos STM em uma base de dados gerada a partir de notícias de jornais digitais, a fim de analisar o que está sendo relatado pela mídia em relação à poluição plástica marinha e como isso pode variar de acordo com o alinhamento político. De maneira

geral, a partir de 36 tópicos gerados foi constatada uma evolução na cobertura midiática em relação a este tema, tendo como ênfase as notícias que explicam os problemas atuais relacionados aos plásticos no ambiente marinho.

Lee et al. (2021) aplicou modelagem de tópicos a fim de analisar as notícias internacionais relacionadas especificamente ao uso de máscara antes e depois da propagação da pandemia, com o intuito de identificar os principais tópicos sobre uso de máscara nas notícias analisadas e como esses tópicos se relacionam. Os tópicos foram gerados a partir do modelo de tópicos STM, sendo utilizado esse modelo por conta dos seus métodos para visualização de dados que possibilitam uma análise mais profunda dos tópicos gerados. Como resultado, a pesquisa evidenciou a evolução de tópicos com temas “Obrigatório o uso de máscaras”, “Surto de COVID-19 na China”, “Gerenciamento de quarentena” com o início da pandemia. Além disso, reafirmou a importância estatística de pesquisas com grande volume de dados nos estudos de ciências sociais.

Teh et al. (2022) em sua pesquisa tiveram como objetivo apresentar uma análise dos tópicos populares nas mídias sociais relacionados à poluição plástica, por meio da aplicação dos algoritmos *Latent Semantic Indexing* (LSI), *Non-Negative Matrix Factorization* (NMF), LDA e HDP. A base de dados utilizada nesta pesquisa foi gerada a partir de 274.404 mil *tweets* coletados da mídia social Twitter. Como resultado, foi possível determinar os principais tópicos relacionados a poluição marinha no Twitter, sendo eles “resíduo zero”, “sustentabilidade” e “conceito 3R”.

3 MATERIAIS E MÉTODOS

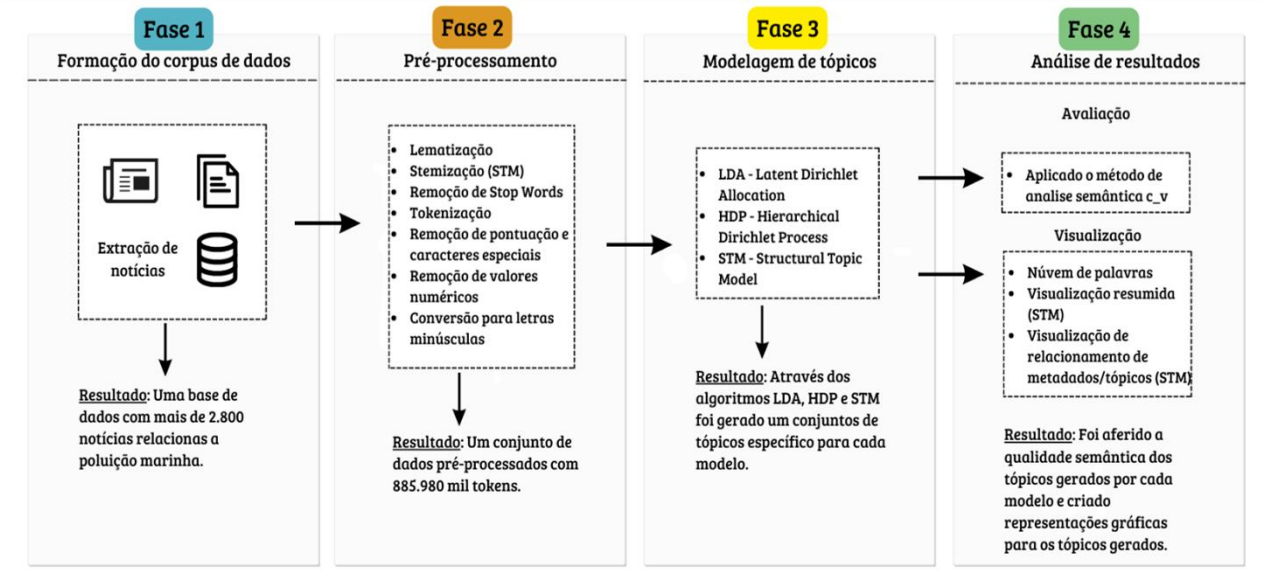
A presente pesquisa possui uma abordagem quali-quantitativa, a parte quantitativa refere-se a identificação dos resultados dos três algoritmos de modelagem de tópicos aplicados por meio da métrica analisada, do ponto de vista qualitativo tem-se a interpretação dos tópicos identificados, por exemplo, na nuvem de palavras.

A caracterização da pesquisa quanto à natureza, é aplicada e de base tecnológica, voltando-se ao uso de algoritmos de aprendizado de máquina no domínio de textos jornalísticos sobre a poluição marinha, disponíveis digitalmente. Em relação aos objetivos é caracterizada como uma pesquisa descritiva, pois emprega métodos para a interpretação dos dados (TRIVIÑOS, 2011), identificando a frequência dos tópicos no *corpus*, sem entrar no mérito de seus conteúdos.

No que se refere aos procedimentos se classifica como bibliográfica e experimental, pois envolve a manipulação de variáveis e a observação dos seus efeitos sobre o objeto de estudo (WAZLAWICK, 2021).

Nesta pesquisa aplicaram-se os algoritmos de modelagem de tópicos LDA, HDP e STM em uma base de dados criada a partir de notícias referentes a poluição marinha disponíveis na internet. Os modelos obtidos por meio da aplicação de cada um dos algoritmos, foram avaliados por meio de processos de análise semântica, a fim de obter dados estatísticos que representam a capacidade de cada modelo em obter tópicos coerentes de acordo com o *corpus* analisado.

Figura 1 — Principais etapas na aplicação de modelagem de tópicos



Fonte. Do autor

A arquitetura geral desta pesquisa (Figura 1), compreende quatro etapas essenciais, iniciando com a extração de notícias relacionadas à poluição marinha publicadas em sites públicos na internet. Após isso, realizou-se a etapa de pré-processamento dos dados a fim de garantir a qualidade dos modelos gerados por meio dos algoritmos de modelagem de tópicos. Posteriormente, executou-se a etapa de modelagem de tópicos por meio da aplicação dos algoritmos LDA, HDP e STM. Por fim, a avaliação dos resultados foi realizada pela análise semântica da qualidade dos tópicos identificados e pela visualização destes por meio das representações gráficas.

3.1 FORMAÇÃO DO CORPUS DE DADOS

Foi desenvolvido nesta etapa um *corpus* de dados formado por textos jornalísticos relacionados a poluição marinha dos anos de 2018, 2020 e 2022 extraídos da internet, para isso aplicou-se o *web scraping* que consiste em uma técnica de coleta automática de dados de sites, páginas web e redes sociais. A metodologia utilizada consistiu na reunião dos dados extraídos em um arquivo *.csv*, por meio de *script* em linguagem de programação Python, versão 3.9.7.

Inicialmente, foram definidos termos de busca para o domínio de aplicação referente a poluição marinha e o período das notícias, a fim de refinar a busca. Os termos utilizados incluíram as seguintes combinações de palavras-chave, (“marinho” OR “oceano” OR “mar” OR “litoral” OR “praia”) AND (“poluição”). O conjunto de anos utilizado na busca das notícias representa três períodos recentes e que foram correlacionados com a pandemia de Covid-19. O ano de 2018 refere-se a um corte de dois anos que antecederam o início da pandemia, o ano de 2020 quando se iniciou a pandemia e dois ano após ao seu início, ano de 2022, em que ainda se vivia o período pandêmico.

A biblioteca Gnews[2], disponível para a linguagem de programação Python,

[2] Disponível em: https://github.com/ranahaani/GNews

foi utilizada para realizar a busca de notícias no Google News, com base nos termos definidos anteriormente. A amostragem dos dados para o *corpus* foi coletada no dia 20 de maio de 2023. A função *gnews.search()* foi empregada e obteve um número de 2.883 textos jornalísticos relacionados a poluição marinha, disponibilizados em portais de notícias do Brasil. Além do corpo de cada texto jornalístico, também foram extraídos o título, a data de publicação, o *link* da notícia e o portal responsável pela publicação. Após a obtenção dos resultados da busca, as notícias foram armazenadas em um arquivo *.csv*.

Os documentos coletados, organizados por corpus que foram utilizados nesta pesquisa estão disponibilizados no GitHub, no link: https://github.com/gustavoPresa/noticias-poluicao-marinha-dataset.

3.2 PRÉ-PROCESSAMENTO DOS DADOS

Após a formação do *corpus* de dados de notícias relacionadas à poluição marinha, realizou-se um conjunto de etapas de pré-processamento essenciais para a posterior execução do PLN por meio dos algoritmos de modelagem de tópicos.

No pré-processamento dos dados foram aplicadas as técnicas de lematização, *stemização*, remoção de *stopwords*, remoção de caracteres especiais e pontuação, remoção de valores numéricos e conversão das palavras para minúsculo.

As técnicas de lematização e stemização foram aplicadas visando reduzir as palavras para suas formas base, sendo a stemização aplicada apenas para a execução do algoritmo STM, isso pois este modelo aplica a stemização de forma padrão ao ser executado. Aplicou-se a lematização nas palavras com o intuito de unificar variações morfológicas, mantendo a forma raiz das palavras. Por exemplo, palavras como "poluição" e "poluindo" reduziram-se ao lema "poluir". A stemização se assemelha a lematização, porém a transformação da palavra ocorre de forma mais simples, consistindo em remover os sufixos das palavras para reduzi-las à sua forma raiz, também conhecida como *stem*. Por exemplo, palavras como "poluição" e

"poluindo" reduziram-se ao *stem* "polui".

No processo de tokenização, os textos foram divididos em *tokens*, que são unidades individuais de palavras ou caracteres, para facilitar o processamento posterior. A diferenciação de palavras com base em letras maiúsculas e minúsculas foi evitada convertendo-se todas para minúsculo, o que garantiu as palavras com a mesma raiz, mas com diferenças de capitalização, fossem tratadas como iguais. Além disso, etapas de remoção de caracteres, de pontuações, de valores numéricos e de palavras muito comuns que não carregam significado relevante para a análise, como por exemplo, artigos, preposições e pronomes também foram aplicadas no conjunto de dados.

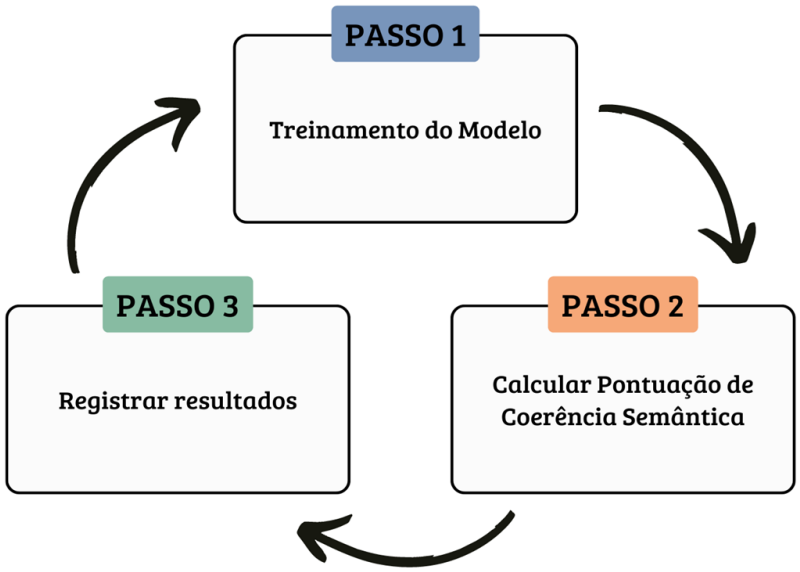
Ao aplicar essas técnicas de pré-processamento na base de dados coletada, buscou-se obter um conjunto de dados mais limpo e padronizado, onde as palavras-chave relevantes para a análise da poluição marinha pudessem ser identificadas de forma mais precisa. Isso facilita a realização de análises estatísticas, a extração de informações e a identificação de padrões nos dados coletados. Como resultado dessa etapa, 885.980 *tokens* foram gerados a partir do *corpus* de dados sobre notícias relacionadas à poluição marinha.

3.3 MODELAGEM DE TÓPICOS

Nesta etapa metodológica da pesquisa realizou-se a execução da modelagem de tópicos por meio dos algoritmos LDA, HDP e STM para analisar um *corpus* de dados de notícias relacionadas a poluição marinha. Na avaliação dos tópicos gerados em cada modelo aplicou-se a métrica de coerência semântica. Além da análise semântica, empregaram-se análises gerais relacionadas aos temas dos tópicos e a visualização dos resultados das modelagens por meio de diferentes gráficos, que foram desenvolvidos a fim de possibilitar uma melhor representação dos tópicos fornecidos por cada algoritmo.

O *corpus* de dados previamente coletado e pré-processado foi submetido aos algoritmos LDA, HDP e STM a fim de extrair uma sequência de tópicos. Nos diferentes treinamentos do modelo, extraiu-se 5, 10, 15 e 20 tópicos com o intuito de analisar a coerência semântica dos modelos com mais de um valor de referência. Os passos efetuados no treinamento dos modelos são apresentados na Figura 2.

Figura 2. Fluxo de treinamento dos modelos de tópicos



3.3.1 Aplicação do algoritmo LDA

O algoritmo de Alocação Latente de Dirichlet (LDA) foi implementado utilizando a biblioteca Gensim[3], disponível na linguagem de programação Python. O algoritmo foi executado com os seguintes parâmetros: *corpus*, que representa o conjunto de documentos pré-processados; *id2word*, que é o dicionário filtrado contendo o mapeamento de palavras para índices; *num_topics*, parâmetro que define o número de tópicos que o modelo deve identificar, este parâmetro é necessário pois

[3] Disponível em: https://radimrehurek.com/gensim/

o LDA é um modelo de tópicos baseado em aprendizado de máquina supervisionado; e *passes*, que indica o número de iterações que o algoritmo deve realizar no *corpus* durante o treinamento.

Os números de tópicos utilizados para treinamento foram os valores presentes na lista de valores previamente definida nesta pesquisa, já o número de passes foi definido por meio da métrica de perplexidade gerada pela função *log_perplexity*, disponível na biblioteca Gensim. Essa métrica fornece uma medida numérica da capacidade do modelo de generalizar e fazer previsões sobre novos documentos.

3.3.2 Aplicação do algoritmo HDP

Na implementação do algoritmo Processo Hierárquico de Dirichlet também se utilizou a biblioteca Gensim por meio da função *HdpModel*. Esta função recebeu como parâmetros o *corpus*, que representa o conjunto de documentos pré-processados, e o dicionário filtrado, *id2word*, contendo o mapeamento de palavras para índices.

Diferentemente do modelo LDA, não foi necessário especificar um número fixo de tópicos a serem gerados ou o número de *passes* a serem executados. Isso ocorre porque o modelo HDP é capaz de inferir automaticamente o número de tópicos a partir dos dados, pois é um método de aprendizado de máquina não supervisionado (DAI; STORKEY, 2015, tradução nossa).

O algoritmo HDP ao não especificar um número fixo de tópicos ou *passes*, se ajusta automaticamente aos dados disponíveis, adaptando-se a complexidade e a variabilidade do conjunto de documentos. Nesta pesquisa, o algoritmo HDP identificou um número total de 149 tópicos, mas para fins de comparação, apenas os resultados dos 5, 10, 15 e 20 primeiros tópicos foram analisados.

3.3.3 Aplicação do algoritmo STM

A implementação do algoritmo Modelo de Tópicos Estrutural foi realizada por meio da linguagem de programação R, utilizando o pacote *stm*. A função *stm()* do pacote permite a criação e o treinamento do modelo, utilizando os seguintes parâmetros, *vocab*, *documents* e *data*. O *vocab* se refere ao vocabulário das palavras contidas nos documentos analisados; o *documents* representa os documentos pré-processados que compõem o *corpus* de análise; o *data* corresponde aos metadados associados aos documentos, que fornecem informações contextuais adicionais.

No contexto deste estudo, os metadados incluem informações como a editora (*publisher*) e o ano de análise (*analyzed_year*). Essas informações podem influenciar a distribuição dos tópicos e auxiliar no entendimento dos resultados obtidos. *Estes metadados foram enviados para o modelo através do parâmetro Prevalence*, que estabelece a estrutura de prevalência do modelo, ou seja, os fatores que influenciam a distribuição dos tópicos nos documentos.

Para o treinamento desse modelo, foi utilizado o parâmetro ~ *publisher* + *s(analyzede_year)*, considerando a editora (*publisher*) e o ano de análise (*analyzed_year*) como variáveis de prevalência. Asim, pode-se investigar como esses fatores afetam a presença e a distribuição dos tópicos nas notícias sobre poluição marinha. O número de tópicos, que foi passado por meio do parâmetro *K* foi alterado entre 5, 10, 15 e 20 visando analisar a coerência dos tópicos gerados conforme os parâmetros informados.

Buscando uma melhor interpretação dos tópicos obtidos pelo modelo STM, uma etapa de rotulagem de tópicos foi aplicada. Esta técnica serve para atribuir rótulos ou palavras descritivas a cada um dos tópicos gerados pelo modelo, ajudando assim a compreender e interpretar os resultados do modelo de tópicos.

3.4 AVALIAÇÃO E VISUALIZAÇÃO DOS RESULTADOS

A avaliação dos tópicos gerados pelos modelos LDA, HDP e STM foi

realizada por meio da métrica de coerência semântica dos tópicos gerados para cada modelo, utilizando-se a ferramenta Palmetto[4]. A coerência semântica, segundo CHAUHAN, SHAH (2021) consiste em medir o grau de conexão semântica entre as palavras que compõem cada tópico, com o objetivo de verificar se as palavras associadas a um tópico representam um tema coerente e significativo. A coerência semântica é geralmente calculada em uma escala que varia de 0 a 1, onde valores mais altos indicam uma maior coerência semântica (CHAUHAN; SHAH, 2021).

No cálculo da coerência semântica de cada tópico foi empregado o método coherence_c_v (*c_v*), que se baseia na comparação das probabilidades de coocorrência de palavras nos documentos. Essa análise permitiu identificar qual modelo e número de tópicos apresentaram os resultados mais coerentes e interpretáveis.

Além da análise quantitativa por meio da coerência semântica, também foi realizada uma análise qualitativa dos tópicos gerados pelos modelos. Os 20 principais tópicos de cada modelo foram examinados manualmente utilizando gráficos de nuvens de palavras, considerando as palavras-chave mais relevantes e a sua interpretabilidade em relação ao domínio do conhecimento referente a poluição marinha. Esta análise qualitativa complementou a avaliação da coerência semântica e forneceu informações adicionais sobre a qualidade dos tópicos gerados.

A visualização dos resultados obtidos ocorreu por meio de representações gráficas dos tópicos gerados pelos modelos, que incluíram a geração de gráficos de frequência de palavras nos tópicos, de distribuição dos tópicos ao longo do tempo e de distribuição dos tópicos por editora, quando aplicável. Estas representações gráficas foram geradas utilizando as bibliotecas Gensim e NLTK[5].

[4] Disponível em: https://palmetto.demos.dice-research.org/
[5] Disponível em: https://www.nltk.org/

4 RESULTADOS E DISCUSSÕES

Os resultados obtidos nesta pesquisa foram divididos em duas etapas, análise quantitativa e análise qualitativa. Por meio da métrica quantitativa de coerência semântica, analisou-se os modelos gerados pelos algoritmos LDA, HDP e STM, conforme o número de tópicos definido, nos algoritmos supervisionados (tabela 1).

Tabela 1 — Valor de coerência semântica por tópicos.

	5 tópicos	**10 tópicos**	**15 tópicos**	**20 tópicos**
LDA	0.1095	0.1128	0.1239	0.1077
HDP	0.1091	0.1261	0.1333	0.1381
STM	0.1576	0.1472	0.1358	0.1413

Fonte: Dados da pesquisa, 2023

No que se refere a coerência semântica, pode-se observar na tabela 1 que os algoritmos analisados apresentaram valores diferentes para cada número de tópico selecionado. O algoritmo LDA apresentou seus valores mais altos com 15 tópicos, mostrando uma tendência de criação de tópicos menos interpretáveis e distintos a partir deste valor. O algoritmo HDP demonstrou uma evolução nos valores de coerência semântica conforme o número de tópicos aumentou, tendo seu melhor resultado, dentro dos parâmetros analisados, com 20 tópicos. O algoritmo STM teve seu maior valor de coerência semântica com 5 tópicos, demonstrando uma tendência de aumento nos valores analisados a partir de 15 tópicos.

Os resultados desta análise mostraram que cada algoritmo alcançou sua melhor pontuação de coerência semântica com diferentes números de tópicos, mesmo quando aplicados ao mesmo *corpus* de dados. Sendo possível evidenciar que a capacidade de geração de modelos semanticamente coerentes varia de acordo com seu contexto de aplicação.

A avaliação qualitativa referente aos modelos utilizados foi gerada a partir de gráficos de nuvens de palavras, que representa as principais palavras encontradas

em um conjunto de tópicos gerados por um modelo de tópicos. Essa representação visual é construída com base na frequência das palavras, ou seja, quanto mais frequente uma palavra é nos tópicos, maior é o seu destaque na nuvem de palavras.

Para fins de comparação, nesta pesquisa as nuvens de palavras foram geradas a partir dos 20 principais tópicos de cada modelo analisado.

Figura 3 — Nuvem de palavras gerado pelo modelo LDA

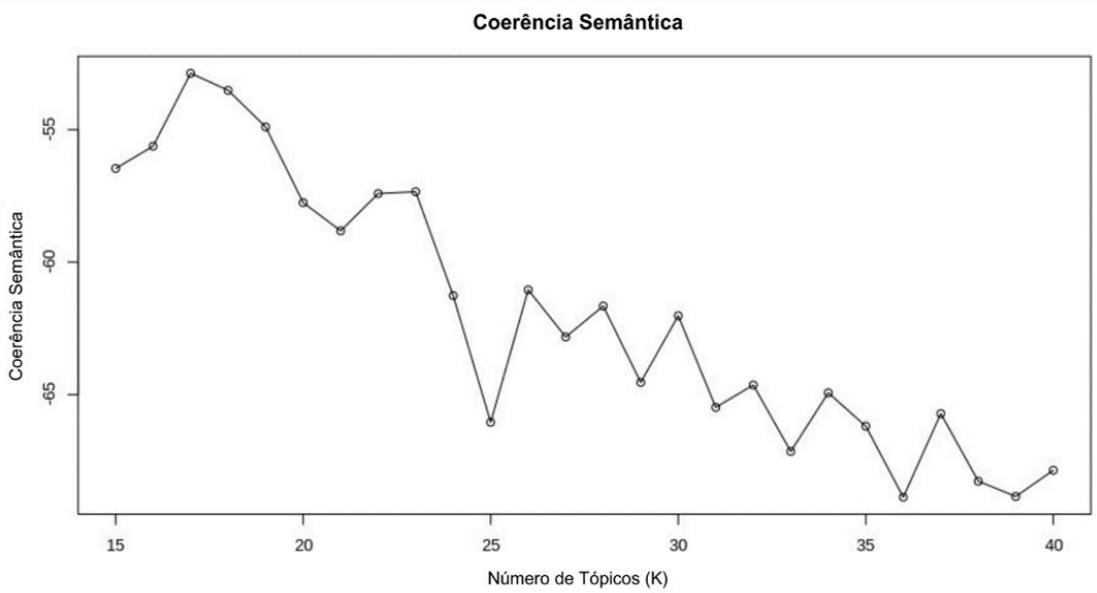
Fonte. Do autor.

Os três modelos obtiveram resultados semelhantes ao identificar as palavras mais importantes dos tópicos, dando ênfase a palavras como, “plástico”, “lixo”, “oceano”, “mar”, entre outras que representam o contexto analisado. Porém,

observa-se que os modelos LDA e HDP obtiveram maior eficiência na redução de ruído e de volume de dados desnecessários durante a análise de texto. De forma geral, os modelos analisados apresentaram bons resultados na visualização e comunicação dos resultados, apresentando de forma concisa e acessível o tema principal que está sendo analisado, neste caso, a poluição marinha.

A análise dos principais tópicos relacionados a poluição marinha discutidos nos textos jornalísticos analisados foi feita por meio do modelo STM por conta da sua capacidade de analisar os metadados associados aos documentos durante o processo de modelagem de tópicos. Nesta etapa, o modelo foi treinado para gerar um total de 17 tópicos. Este valor foi definido por meio da função *searchK* do pacote *stm*. Esta função calcula a coerência semântica para cada valor de tópico e retorna um gráfico com os resultados, como pode-se observar na figura 6. Este gráfico apresenta o valor de 17 tópicos como sendo o ideal para o modelo STM em termo de coerência semântica. Considerando-se as diferentes abordagens técnicas utilizadas pela função searchK comparado a ferramenta Palmetto para análise de coerência semântica, nesta pesquisa os resultados obtidos pela função searchK foram desconsiderados na avaliação quantitativas de coerência semântica, servindo apenas como auxiliar na análise qualitativa.

Figura 6 — Gráfico de análise de números de tópicos



Fonte. Do autor.

Os resultados obtidos pelo modelo STM após a coleta dos 17 principais tópicos e rotulagem dos mesmos é apresentada na Tabela 2.

Tabela 2. Dados da rotulagem de tópicos.

N.º do Tópico	Palavras do Tópico	Tipo de poluição	Rótulo
Tópico 2	oceano ambient plástico poluição água email ecodeb	Ambiental	Poluição por lixo
Tópico 3	ambient caso público água empresa esgoto poluição	Ambiental	Poluição por esgoto
Tópico 4	espéci plástico peix animai marinha marinho encontra	Fauna	Poluição por lixo
Tópico 5	ilha navio praia ambient turista marinha brasil	Ambiental	Poluição por óleo
Tópico 6	rio mar ambient projeto brasil oceano água	Generalista	Projeto Ambiental
Tópicos 8	plástico lixo oceano animai encontrar marinho marinha	Ambiental/Fauna	Poluição por lixo
Tópico 9	óleo mancha praia ambient petróleo litor mar	Ambiental	Poluição por óleo
Tópico 10	praia água rio esgoto mar banho ambient	Ambiental	Poluição por esgoto

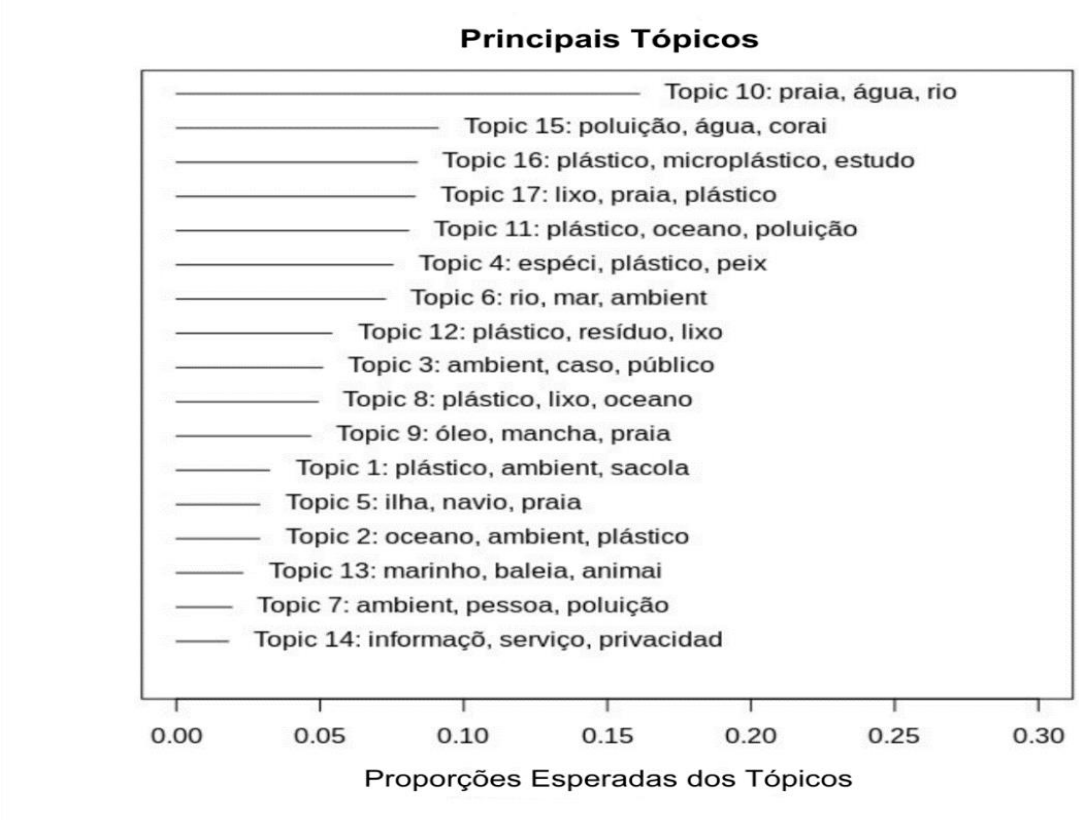
Tópico 11	plástico oceano poluição ambient mundo lixo país	Ambiental	Poluição por lixo
Tópico 12	plástico resíduo lixo estudo brasil tonelada oceano	Ambiental	Poluição por lixo
Tópico 13	marinho baleia animai ruído navio marinha poluição	Fauna	Poluição sonora
Tópico 15	poluição água corai emissõ ambient mudança aumento	Fauna	Mudanças Climáticas
Tópico 16	plástico microplástico estudo oceano partícula produto mar	Ambiental	Poluição por lixo
Tópico 17	lixo praia plástico resíduo mar projeto ambient	Ambiental	Poluição por lixo

Fonte: Dados da pesquisa, 2023.

Dos 17 tópicos gerados pelo modelo, 14 deles puderam ser devidamente associados e rotulados com temas sobre poluição marinha. Dentre eles, 7 tópicos foram rotulados como “Poluição por lixo”, 2 tópicos como “Poluição por Esgoto”, 2 tópicos como “Poluição por Óleo” e “Poluição Sonora”, “Projeto Ambiental” e “Mudanças Climáticas” tiveram um tópico cada. Os resultados desta análise mostraram a poluição por lixo plástico como o problema de poluição marinha mais evidente em textos jornalísticos veiculados nacionalmente.

A popularidade de cada tópico gerado pelo modelo STM pode ser representada por meio da figura 8, onde os 17 tópicos gerados estão ordenados por sua prevalência, permitindo visualizar a relevância de cada tópico em relação aos outros.

Figura 7 — Principais tópicos gerados pelo modelo STM

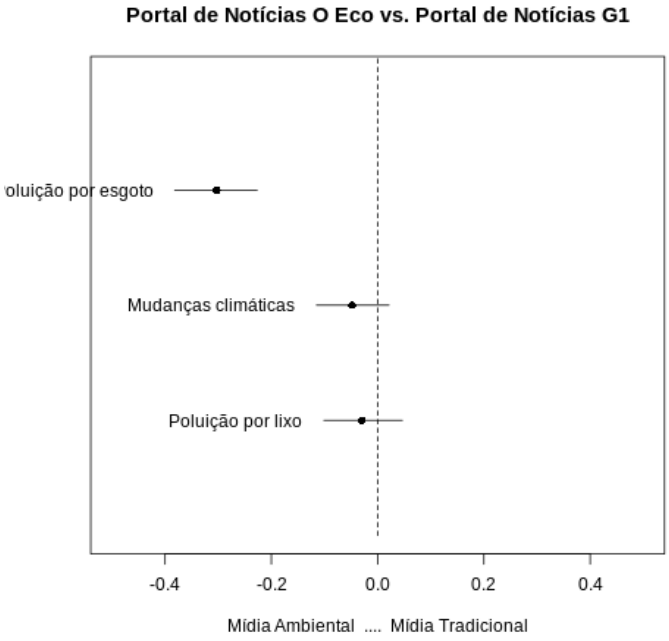


Fonte. Do autor.

Referindo-se a Figura 7, as notícias analisadas extraídas nas datas de 2018, 2020 e 2022 apresentam como principais temas, respectivamente, “Poluição por Esgoto”, “Mudanças Climáticas” e “Poluição por Lixo”, sendo estes os três tópicos predominantes nos documentos analisados.

Com intuito de analisar a divulgação de textos jornalísticos relacionados a poluição marinha em diferentes portais de notícias, um gráfico comparativo (figura 9) entre os portais de notícias “O eco” e “G1” foi gerado. O portal de notícias “O eco” atua diretamente no jornalismo ambiental, já o portal de notícias “G1”, representa um grande portal de notícias nacional.

Figura 8 — Gráfico de comparação de tópicos entre mídias



Fonte. Do autor.

Os resultados obtidos na comparação acima apresentam uma influência preponderante da mídia ambiental “O Eco” em relação a mídia tradicional “G1” em na divulgação de notícias sobre os três principais tópicos identificados pelo modelo STM. Evidenciando assim a importância de tais veículos de imprensa na divulgação de notícias sobre poluição marinha em âmbito nacional.

5 CONCLUSÃO

Este trabalho teve como objetivo avaliar quantitativamente a modelagem automática de tópicos em textos jornalísticos sobre poluição marinha nos anos de 2018, 2020 e 2022. Também, buscou-se demonstrar a capacidade de tais tecnologias e métodos para fins científicos, explorando os modelos de tópicos para identificar e analisar os principais temas relacionados à poluição marinha discutidos nos textos jornalísticos nacionais.

A análise quantitativa foi realizada por meio do método de avaliação de coerência semântica. Essas diferenças entre os modelos *Latent Dirichlet Allocation*

(LDA), *Hierarchical Dirichlet Process* (HDP) e *Structural Topic Model* (STM) se refletiram nos valores de coerência semântica obtidos, mesmo quando os modelos foram aplicados à mesma base de dados. Os resultados apresentados nesta etapa indicam que a escolha do algoritmo de modelagem de tópicos pode influenciar diretamente nos resultados e na interpretação dos tópicos gerados, demonstrando ser essencial considerar essas diferenças ao selecionar e avaliar os modelos de tópicos para garantir uma análise precisa e consistente dos dados.

A avaliação qualitativa dos modelos apresentou resultados consistentes ao identificar palavras-chave como "plástico", "lixo", "oceano" e "mar" nos gráficos de nuvens de palavras, sendo relevantes para o contexto da poluição marinha. Forneceram também uma visualização clara e acessível dos principais tópicos abordados. Os modelos LDA e HDP foram mais eficientes na redução de ruído e dados desnecessários durante a análise textual.

O modelo STM apresentou bons resultados ao gerar gráficos que correlacionam os tópicos com os metadados relacionados ao mesmo. Porém, a falta de eficiência computacional comprometeu a passagem dos parâmetros "prevalence" e "content" no treinamento do modelo, o que implicou na capacidade de capturar adequadamente a estrutura dos tópicos e a distribuição de palavras relevantes em cada tópico, gerando resultados menos precisos e representações menos significativas dos tópicos.

Em pesquisas futuras recomenda-se a aplicação dos modelos de tópicos em um ambiente computacional de alto desempenho, possibilitando mais testes relacionados os parâmetros de treinamento de cada modelo. Mais representações gráficas dos tópicos gerados também podem ser aplicadas a fim de obter novos resultados. Além disso, a análise aplicada em um período maior de tempo pode auxiliar na comparação da evolução de tópicos dentre os anos.

Referências

BLEI, D. M. Probabilistic topic models. Communications of the ACM, New York,

v. 55, n. 4, p. 77–84, 2012.

CHAUHAN, Uttam; SHAH, Apurva. Improving Semantic Coherence of Gujarati Text Topic Model Using Inflectional Forms Reduction and Single-letter Words Removal. **Acm Transactions On Asian And Low-Resource Language Information Processing**, [S.L.], v. 20, n. 1, p. 1-18, 31 jan. 2021. Association for Computing Machinery (ACM). http://dx.doi.org/10.1145/3447760.

FORLEO, M. B, ROMAGNOLI, L. Marine plastic litter: public perceptions and opinions in Italy. 2021. Pollut. Bull.165
p. 112160.

GONZALEZ, M.; LIMA, V. L. S. Recuperação de informação e processamento da linguagem natural. In: XXIII Congresso da Sociedade Brasileira de Computação, 2003. p. 347-395.

KELLER, Ellis; WYLES, Kayleigh J.. Straws, seals, and supermarkets: topics in the newspaper coverage of marine plastic pollution. Marine Pollution Bulletin, [S.L.], v. 166, p. 112211, maio 2021. Elsevier BV.

LEE, K.R. et al. Intuitive Topic Structural Topic Model Analysis of Mask-Wearing Issue Using International News Big Data. *Int. J. Environ. Int. J. Environ. Res. Public Health* 2021, *18*, 6432.

MARQUESONE, R. Big data: técnicas e tecnologias para extração de valor dos dados. São Paulo: Casa do Código, 2016.

OTERO, P.; GAGO, J.; QUINTAS, P.. Twitter data analysis to assess the interest of citizens on the impact of marine plastic pollution. Marine Pollution Bulletin, [S.L.], v. 170, p. 112620, set. 2021. Elsevier BV.

ROBERTS, Margaret e *et al*. The structural topic model and applied social science. Advances In Neural Information Processing Systems Workshop On Topic Models: Computation, Application, And Evaluation, [*s. l*], v. 4, p. 1-20, 10 dez. 2013.

TEH, Phoey L., SCOTT PIAO, Mansour Almansour, HUEY F. Ong, and Abdul Ahad. 2022. "Analysis of Popular Social Media Topics Regarding Plastic Pollution" Sustainability 14, no. 3: 1709.

TRIVIÑOS, Augusto Nibaldo Silva. Introdução à pesquisa em ciências sociais: a pesquisa qualitativa em educação. São Paulo: Atlas, 2011.

ZEROUAL, Imad; LAKHOUAJA, Abdelhak. Data science in light of natural language processing: an overview. Procedia Computer Science, [S. l.], v. 127, p. 82–91, 2018.

ZHU, D. et al. Intuitive Topic Discovery by Incorporating Word-Pair’s Connection Into LDA. In: 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and

Intelligent Agent Technology, p. 303-310, 2012.

WAZLAWICK, Raul. Metodologia de pesquisa para ciência da computação. Rio de Janeiro: LTC, 2021.