

ARTIGO

Como estudar mídias digitais em tempos de dilúvio informacional? O protocolo metodológico TUCANO aplicado à pesquisa documental em notícias online

Victor Felipe Mendanha de Lima 
e Frederico Marianetti Soriani 

Resumo

Este artigo apresenta o TUCANO (Técnicas Unificadas de Categorização e Análise de Notícias Online), protocolo metodológico híbrido para análise de corpora de notícias online em português. Desenvolvido para enfrentar um desafio das amostragens não probabilísticas em Ciências Sociais, o TUCANO busca delimitar, organizar e explorar conjuntos de notícias sem perder a densidade interpretativa da análise qualitativa. O protocolo articula técnicas de NLP, como tokenização, vetorização e análise de coocorrência, com uma lógica teórica orientada à construção de corpora e à subamostragem temática. Seu fluxo de trabalho organiza-se em seis etapas, da formulação da pergunta-guia à definição de recortes fundamentados para análise qualitativa. A aplicabilidade é demonstrada em um estudo sobre a cobertura midiática do uso de canudos plásticos e seus impactos na vida marinha. Com uma infraestrutura computacional leve, transparente e acessível, o TUCANO potencializa a análise qualitativa e amplia o uso crítico de ferramentas computacionais nas Ciências Sociais aplicadas.

Palavras-chave

Estudos sociais da ciência e da tecnologia; Profissionalismo, desenvolvimento profissional e formação em divulgação científica; Ciência e mídia

Recebido em: 19 de Março de 2026

Revised: 25 de Maio de 2026

Aceito em: 24 de Julho de 2026

Publicado em: 21 de Setembro de 2026

1 ▪ Introdução

1.1 ▪ *Jornalismo, mediação e tensões contemporâneas*

À medida que estratégias de mercado passaram a ocupar o lugar de valores éticos e de causas públicas, a imprensa foi se consolidando como um produto industrial, enquanto a burguesia ascendente passou a ocupar o espaço público e a estabilizar suas representações da realidade [Pena, 2021]. Os modos de narrar o mundo são intrinsecamente reagentes a disputas por autoridade, legitimidade e diálogo dos afetos [Schudson, 1981; Gould & Rocha, 2021].

O jornalismo é um espaço de coexistência entre permanências e rupturas, no qual lógicas tradicionais de mediação dialogam com sensibilidades emergentes, afetos e novas formas de visibilidade. Parte dessas transformações concebe o social como algo construído nas relações entre os actantes e nas dinâmicas de força que os conectam [Medina, 2002; van Dijck, 2013; Gould & Rocha, 2021].

A ecologia comunicacional digital intensificou esse cenário, ampliando o ritmo e a complexidade dessas disputas. O dilúvio informacional, as infraestruturas algorítmicas e a proliferação de plataformas digitais criam ambientes nos quais a notícia não apenas circula, mas se integra a fluxos híbridos, combinando traços humanos, automações e lógicas de viralização [Castells, 2009; Santos De Mattos, 2023; Hissa, 2024; Souza & Coutinho, 2025]. A crescente porosidade das fronteiras entre jornalismo, entretenimento, ciência e política dificulta a compreensão de como determinados temas adquirem centralidade e se estabilizam no debate público [Walsh-Childers & McKinnon, 2024].

Esse cenário impõe um desafio metodológico. A análise da mídia digital exige ferramentas capazes de lidar com grandes volumes de informação sem perder de vista as dimensões discursivas, sociais e simbólicas que estruturam a produção de sentidos na comunicação pública [Castelfranchi, 2017].

1.2 ▪ *Tensões epistemológicas e desafios metodológicos nas Ciências Sociais*

As Ciências Sociais desenvolveram um arsenal metodológico robusto voltado à compreensão de símbolos complexos. Abordagens como, por exemplo, a Análise de Conteúdo, as Semióticas e a Análise de *Framing*, permitem acessar significados profundos e reconstruir processos interpretativos, evitando reduções simplificadoras de fenômenos multifatoriais [Gil, 2024]. A flexibilidade metodológica constitui uma das principais forças dessas abordagens em Ciências Sociais, pois possibilita ao pesquisador selecionar os instrumentos adequados ao problema investigado. Essa flexibilidade não busca esgotar a realidade empírica, mas torná-la inteligível dentro de uma moldura interpretativa cuidadosamente construída [Gil, 2024].

Essa mesma abertura encontra limites práticos quando confrontada com campos teóricos próprios bem delimitados e, por vezes, antagônicos, como no debate entre métodos quantitativos e qualitativos [Nunes, 2007]. Análises manuais qualitativas, embora minuciosas, são inevitavelmente lentas, dependem de recortes reduzidos e lidam com a subjetividade inerente às escolhas do pesquisador [Nunes, 2007; Gil, 2024]. Bauer e Gaskell [2017] destacam que a própria decisão sobre o que analisar já estabelece um horizonte

interpretativo restrito, capaz de obscurecer padrões discursivos que só emergem em larga escala. Essa limitação não configura um erro metodológico, mas uma característica constitutiva da abordagem qualitativa [Nunes, 2007]. O desafio é mais premente em ecossistemas informacionais densos, nos quais a quantidade de dados excede a capacidade de análise humana tradicional.

1.3 ■ *A virada computacional: promessas, riscos e críticas*

O aumento do volume de dados e da complexidade informacional impulsionou a adoção de métodos computacionais para a análise de grandes *corpora* textuais (e.g., Blei et al. [2003], Grimmer e Stewart [2013], Castelfranchi [2017], Moreira et al. [2022], van Loon [2022] e Neves e Massarani [2022]). Segundo Castelfranchi [2017], esse cenário criou um verdadeiro “Laboratório a céu aberto” para as Ciências Sociais, no qual a disponibilidade massiva de documentos, em formato digital, pode ser analisada em larga escala para compreender padrões matemáticos.

Essa ampliação do alcance analítico computacional não ocorre sem riscos. Boyd e Crawford [2012] alertam para a adoção acrítica dessas ferramentas, que pode gerar uma confiança excessiva em padrões estatísticos e a suposição equivocada de que algoritmos são capazes de substituir a interpretação humana. Castelfranchi [2017] ressalta que a abundância de dados não substitui a necessidade de interpretação teórica, uma vez que identificar padrões estatísticos não equivale, por si só, a explicar fenômenos sociais ou compreender seus sentidos atribuídos na sociedade. Métodos computacionais devem ser compreendidos como instrumentos analíticos, e não como instâncias interpretativas finais [Grimmer & Stewart, 2013]. Longe de representar um obstáculo, essa tensão constitui um terreno epistemologicamente fértil para o desenvolvimento de fluxos de trabalho que articulem técnica e teoria.

É dentro desse cenário que se torna pertinente conceber métodos de pesquisa documental que permitam articular técnicas computacionais automatizadas de análise de notícias e decisões analíticas teoricamente orientadas que subsidiem processos de subamostragem para análises qualitativas. O desafio consiste em estruturar um fluxo de trabalho que seja acessível, transparente e consistente, capaz de ampliar a sensibilidade teórica do pesquisador através de contínua interação com os dados empíricos [Glaser, 1978].

Nosso objetivo é apresentar o TUCANO, um protocolo metodológico híbrido (quanti-qualitativo) para pesquisa documental, em larga escala, de notícias online, em língua portuguesa, baseado em ferramentas e técnicas de processamento de linguagem natural já existentes. O TUCANO busca articular a exploração automatizada de grandes volumes textuais e a subamostragem de *corpus* analíticos teoricamente informadas. E, com o objetivo de elucidar seu uso, o protocolo metodológico foi aplicado em um estudo de notícias sobre o uso de canudos plásticos e a conservação da vida marinha. Nosso objetivo é explorar, de maneira sistemática e ilustrativa, os diferentes níveis (e liberdades) analíticos previstos no fluxo de trabalho do TUCANO.

2 - Protocolo proposto: TUCANO

Abordagens de pesquisa documental de natureza qualitativa tradicionalmente estudam amostras selecionadas por intencionalidade. Em contraste, abordagens de natureza quantitativa operam com *corpora* amplos, mas correm o risco de reificar padrões estatísticos sem uma articulação teórica consistente [Nunes, 2007; Sampaio et al., 2022]. Essa dicotomia entre profundidade interpretativa e escala empírica evidencia uma lacuna metodológica central: como articular volume, rigor e sensibilidade analítica em um fluxo de trabalho acessível?

É nesse interstício que se insere o protocolo metodológico aqui proposto. O TUCANO — **T**écnicas **U**nificadas de **C**ategorização e **A**nálise de **N**otícias **O**nline — foi concebido como um protocolo de trabalho híbrido, capaz de transitar entre a macroanálise quantitativa e a microanálise qualitativa, preservando a riqueza discursiva do material sem renunciar à capacidade de processamento em larga escala. Sua arquitetura busca viabilizar esse movimento sem exigir competências avançadas em programação, tornando o protocolo metodológico acessível.

A proposta fundamenta-se na materialização da pergunta de pesquisa em um dispositivo de escalonamento lexical. A construção de combinações de palavras-chave (*query string*), organizadas sistematicamente (definidas na seção Etapas do TUCANO), permite segmentar o objeto empírico de forma controlada. Esse procedimento orienta a coleta de dados nos motores de busca de notícias online, resultando na constituição de *corpora* amplos e documentados.

A esse fluxo são combinadas etapas de processamento de linguagem natural (NLP), que não apenas revelam padrões, mas também funcionam como um filtro crítico para a etapa qualitativa. Dessa forma, as análises interpretativas subsequentes deixam de depender exclusivamente de amostras por intencionalidade e passam a apoiar-se em uma exploração prévia do *corpus*, assegurando maior representatividade na seleção dos textos.

O objetivo do TUCANO não é substituir a análise qualitativa, mas criar as condições de possibilidade para o uso instrumental da análise computacional em pesquisas documentais. O protocolo metodológico opera como uma ponte entre os métodos tradicionais das Ciências Sociais aplicadas e as abordagens computacionais contemporâneas. O propósito deste artigo é apresentar os fundamentos, o desenho e o potencial analítico do TUCANO em pesquisas documentais em mídias digitais (o que inclui o jornalismo online).

2.1 ■ Fundamentos e gênese do TUCANO

O TUCANO emerge entre o rigor interpretativo das Ciências Sociais e o percurso consolidado do NLP. De um lado, a tradição da pesquisa documental em Comunicação fornece o arcabouço teórico para a interpretação de fenômenos midiáticos. De outro, o NLP oferece um conjunto de técnicas relativamente estáveis e escaláveis capazes de estruturar grandes volumes de textos de forma sistemática [Jurafsky & Martin, 2026].

Essas operações não são neutras. Elas organizam o texto segundo lógicas específicas, que precisam ser compreendidas no processo interpretativo. O TUCANO apropria-se desses métodos, adaptando-os à lógica dos estudos de notícias online. Em vez de buscar interpretações automáticas, o protocolo metodológico prioriza a construção de um fluxo de

preparação de dados transparente e replicável, que sirva de entrada para análises discursivas.

O raciocínio lógico do TUCANO é organizado em seis etapas interligadas (ver Figura 1):

ETAPA 1 - materializar a pergunta da pesquisa em um escalonamento lexical estruturado por palavras-chave (*query string*), que passa a orientar de forma sistemática a coleta do conjunto de documentos (notícias).

ETAPA 2 - coletar as notícias por meio da Plataforma online *Media Cloud* [Roberts et al., 2021]. O objetivo é reunir um conjunto de dados amplo e bem documentado, assegurando tanto a rastreabilidade quanto a replicabilidade do procedimento.

ETAPA 3 - aplicação das técnicas básicas de NLP, voltadas à padronização textual e à redução de ruídos.

ETAPA 4 - representar os títulos em matrizes vetoriais, permitindo a identificação de termos frequentes e suas relações.

ETAPA 5 - produzir visualizações que evidenciam redes ou associações estatísticas, resultando em uma cartografia do *corpus* de análise.

ETAPA 6 - definir critérios que orientam a geração de subconjuntos de documentos destinados à análise qualitativa (*close reading*).

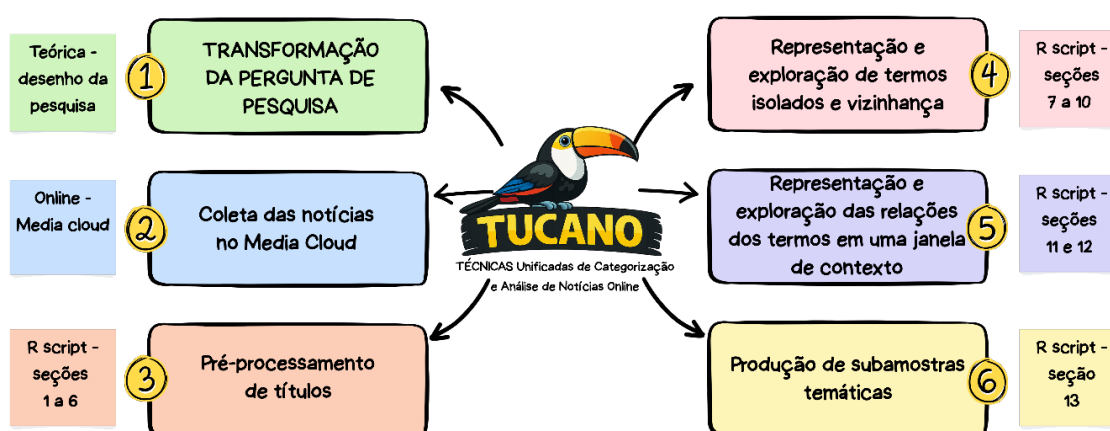


Figura 1. Fluxo de trabalho do TUCANO — Técnicas Unificadas de Categorização e Análise de Notícias Online. FONTE: Elaborado pelos autores, 2026.

As etapas 1 e 2 são etapas que envolvem o processo de formulação do problema e a operacionalização da coleta, e são realizadas fora do ambiente de programação computacional. Já as etapas de 3 a 6 são realizadas no *software* R [R Core Team, 2025], escolhido por ser *open-source* e amplamente conhecido. É indicado ter instalado a versão mais recente do *software*, e indicamos a instalação da interface gráfica RStudio [Posit Team, 2025]. O *script* para análise no R [R Core Team, 2025], *R script*, foi desenvolvido para execução em computadores de uso pessoal, dispensando o uso de servidores dedicados à computação de alto desempenho.

Cada etapa é extensivamente comentada no *R script*, mas, como material complementar, disponibiliza-se um chat (<https://doi.org/10.5281/zenodo.18332942>) baseado em Inteligência Artificial, implementado no NotebookLM (Google), com a finalidade de auxiliar os usuários no detalhamento operacional das etapas e na compreensão de seu encadeamento. Este recurso é alimentado exclusivamente por fontes *offline* oficiais do projeto e deve ser compreendido como uma ferramenta experimental de consulta, sujeita a limitações e imprecisões. Em nenhuma circunstância substitui a descrição formal apresentada neste artigo, que permanece como a fonte primária, controlada e validada do TUCANO.

2.2 ■ *Dos fundamentos de NLP incorporados ao TUCANO*

O *R script* implementa algumas etapas centrais amplamente descritas na literatura da Ciência da Informação e da Linguística Computacional [Jurafsky & Martin, 2026].

No desenvolvimento do TUCANO, optou-se por não empregar técnicas de redução das palavras à sua forma base, como *stemming* e lematização. Essas técnicas podem apagar nuances relevantes em estudos de comunicação, nos quais escolhas lexicais carregam peso discursivo. Preservar a forma original dos termos facilita a interpretação dos resultados, mantém marcas enunciativas e simplifica o processo de filtragem pós-análise (etapa 6). Assim, sacrificou-se certa compactação lexical em favor da preservação da textura discursiva.

2.3 ■ *Etapas do TUCANO*

Ainda que o TUCANO utilize pacotes e técnicas difundidas na literatura do NLP, nossa inovação reside na forma como esses elementos são articulados em um fluxo de trabalho estruturado em seis etapas. Nosso protocolo metodológico propõe um protocolo simples e replicável, que integra etapas de preparação lógica da pergunta de pesquisa, processamento computacional e abertura para análise qualitativa, oferecendo uma sistematização que facilita sua aplicação em diferentes contextos de pesquisa.

2.3.1 ■ *Da formulação da pergunta e do escalonamento da busca*

A primeira etapa do TUCANO operacionaliza a pergunta de pesquisa em um modelo de palavras-chave, a *query string* [Manning et al., 2008]. O protocolo metodológico propõe um escalonamento em três níveis, para delimitar o campo empírico de forma precisa, sem comprometer sua abrangência, a partir de um Guia de coleta que delimite o conjunto de notícias que o pesquisador busca analisar. São os três níveis:

- Núcleo duro: Pares de palavras-chave diretamente relacionados ao fenômeno em estudo. São os pilares conceituais da investigação, assegurando que a coleta permaneça ancorada no objeto da pesquisa.
- Expansão controlada: Variações morfológicas, sinônimos e termos proximais, ampliando o alcance da busca sem provocar deslocamentos temáticos indesejados. Confere flexibilidade à coleta e permite captar variações relevantes.
- Borda semântica: Termos contextuais e adjacentes, capazes de capturar enquadramentos mais amplos e o entorno discursivo do fenômeno investigado. Possibilita articular o tema central a debates mais extensos.

Esse escalonamento dialoga com as discussões sobre a construção de *corpora* de Bauer e Gaskell [2017] e Gil [2024]. Ele estabelece um campo de pertinência bem definido, ao mesmo tempo em que preserva a abertura necessária para captar variações. A partir dessa estrutura, as palavras-chave materializam a pergunta de pesquisa em um formato executável no motor de busca do *Media Cloud*.

Nesta etapa, definem-se o recorte temporal da pesquisa e o acervo de documentos (notícias). Esse procedimento ancora a coleta de dados em critérios explícitos, conforme discutido por Gil [2024].

2.3.2 ▪ *Da coleta das notícias*

A coleta das notícias é realizada no *Media Cloud* (<http://www.mediacloud.org>) [Roberts et al., 2021], uma plataforma *open-source* mantida por um consórcio entre a *University of Massachusetts Amherst*, a *Northeastern University* e o *Media Ecosystems Analysis Group*. O *Media Cloud* agrega vastos acervos midiáticos (jornalísticos ou não) e permite a execução de buscas complexas por expressões booleanas [Roberts et al., 2021]. As consultas são construídas com base no escalonamento definido na etapa anterior e podem ser executadas em diferentes coleções de notícias. Uma característica central do TUCANO é que todas as análises são realizadas a partir dos títulos das notícias. Esta é uma limitação do *Media Cloud* e é abordada no último tópico deste artigo.

Cabe observar que o protocolo metodológico foi concebido e testado especificamente para a análise de títulos de notícias, devido a essa limitação do *Media Cloud*. Mas o pesquisador pode, de forma manual, substituir o conteúdo da coluna *title*, do arquivo .csv, pelo texto integral das matérias, forçando o *R script* a operar com unidades textuais diferentes. Essa é uma possibilidade técnica não validada sistematicamente, devendo ser utilizada com cautela.

2.3.3 ▪ *Do script computacional e liberdades analíticas*

O *R script* possui 13 seções. Todas as etapas são comentadas e seguem a seguinte sequência lógica:

- Importação do arquivo .csv: importação do arquivo .csv exportado do *Media Cloud*, contendo os metadados das notícias. Não é necessário editar o arquivo antes de usar.
- Filtragem mínima: filtragem inicial para remover duplicatas óbvias (títulos idênticos) e verificar a consistência do idioma.
- Criação do *corpus* de análise: títulos das notícias são extraídos e organizados em um *corpus* de análise.
- Limpeza e padronização: tokenização e normalização, que dividem o texto bruto em partes menores e padronizam essas unidades, transformando texto em dados.
- Construção da DFM (*Bag-of-Words*): Matriz Termo-Documento representando a ocorrência de cada termo (*token* normalizado) nos títulos.
- N-gramas: frequências dos termos isolados e de expressões compostas.
- Geração da FCM e cálculo de correlações lexicais: Matriz de Coocorrência de Frequência (FCM) para identificação das associações contextualizadas entre os termos.

- Produção de visualizações exploratórias: nuvens de palavras e redes de coocorrência, que auxiliam na exploração inicial dos padrões discursivos.
- Exportação de resultados: tabelas de resultados exportadas para planilhas, permitindo análises subsequentes em outros *softwares* ou aprofundamento da análise.
- Subamostragem: definir um conjunto de termos para subamostrar (dentro do *corpus* geral) notícias que o pesquisador tiver interesse em analisar qualitativamente.

Além da flexibilidade do *R script*, o TUCANO oferece dois níveis de liberdade analítica após a coleta inicial dos dados, eliminando a necessidade de repetir coletas. O pesquisador pode optar por:

- Analisar o acervo geral — agrupa todas as notícias, e todos os veículos, em um único acervo, tratando o conjunto como um grande campo unificado. Este modo de análise é útil para identificar tendências gerais, mapear termos dominantes ou visualizar redes semânticas mais amplas que perpassam todo o material coletado, oferecendo uma visão macro do fenômeno. Essa estratégia exige cautela analítica, mais descrita abaixo.
- Analisar por veículo ou conjunto de veículos — filtra as notícias por veículo, permitindo a seleção de um específico, ou um subconjunto (como jornais de grande circulação; jornais com vertente editorial conservadora ou progressista; veículos midiáticos generalistas, incluindo publicação de *releases*) ou, ainda, subconjuntos contrastantes (como veículos comerciais *versus* independentes; nacionais *versus* regionais). Essa possibilidade é relevante à luz das discussões de Hallin e Mancini [2004] e Leal et al. [2014], que demonstram que distintos modelos de imprensa produzem práticas editoriais e narrativas específicas. Além disso, o pesquisador pode, após a coleta, filtrar o conjunto de dados, no *R script*, a fim de construir um *corpus* que enquadre estritamente espaços jornalísticos para além das mídias generalistas. O TUCANO permite observar essas diferenças por meio da textura lexical da cobertura, comparando, por exemplo, veículos de diferentes espectros ideológicos.

Em todos os casos, as análises são realizadas diretamente sobre a base de dados já coletada, no *R script*, o que confere agilidade ao processo e otimiza o fluxo de trabalho.

3 - Resultado da aplicação modelo do TUCANO

Cabe ressaltar que a presente aplicação não tem como finalidade discutir os achados sob uma perspectiva teórico-discursiva. A questão investigada é complexa e densa e, para esse exercício metodológico, algumas decisões — como a seleção de palavras-chave na Etapa 1 — foram intencionalmente simplificadas. O conjunto de dados analisado não esgota a profundidade da problemática proposta. A seguir, apresenta-se a aplicação do TUCANO, organizada por etapas, articulando as decisões metodológicas, resultados intermediários, discussões e apontamentos.

3.1 - Etapa 1

- *Pergunta da pesquisa*
“Como mídias brasileiras noticiaram a relação entre o uso de canudos plásticos, poluição marinha e impactos sobre a vida das tartarugas marinhas?”

■ *Guia da coleta*

Notícias de sites de mídias digitais, em língua portuguesa, sobre a relação entre o uso de plásticos de uso único, a poluição dos mares e oceanos e a conservação da vida marinha.

■ *Escalonamento lexical*

A definição e pareamento das palavras-chave, Tabela 1, seguiram rigorosamente a lógica de núcleo duro, expansão controlada e borda semântica.

Tabela 1. Plano cartesiano (5x5) de pareamento de palavras-chave para coleta das notícias, com classificação dos pares segundo Núcleo duro (ND), Expansão controlada (EC) e Borda semântica (BS). Fonte: Elaborado pelos autores, 2026.

<i>Palavras-chave</i>	<i>canudo</i>	<i>canudos</i>	<i>lixo plástico</i>	<i>poluição</i>	<i>microplásticos</i>
<i>tartaruga marinha</i>	ND	ND	ND	EC	EC
<i>tartarugas marinhas</i>	ND	ND	ND	EC	EC
<i>fauna marinha</i>	EC	EC	EC	BS	BS
<i>vida marinha</i>	EC	EC	EC	BS	BS
<i>animais marinhos</i>	EC	EC	EC	BS	BS

■ *Recorte temporal*

O recorte temporal adotado para a coleta compreende o período de 2016 a 2022. Este intervalo foi selecionado por coincidir com o debate público sobre plásticos de uso único no Brasil. O período engloba marcos regulatórios relevantes, como a Lei Estadual do Rio de Janeiro n° 8006/2018 [Assembleia Legislativa do Estado do Rio de Janeiro, 2018], a Lei Municipal de São Paulo n° 17123/2019 [Assembleia Legislativa do Estado de São Paulo, 2019] e a tramitação do Projeto de Lei do Senado Federal n° 92/2018 [Senado Federal do Brasil, 2018], que nacionalizou a discussão sobre a restrição de plásticos descartáveis.

Para além do legislativo, esse intervalo corresponde à consolidação das tartarugas marinhas como “espécies-bandeira” no imaginário coletivo. Tal processo foi impulsionado pela viralização global do vídeo registrado pela pesquisadora Figgner [2015], que documenta a remoção de um canudo da narina de uma tartaruga marinha. Como parte do procedimento desta etapa do estudo, foi realizada, pelos autores, uma reconstrução da circulação histórica do conteúdo do vídeo da Figgner [2015] por meio do *Wayback Machine* (*Internet Archive* — <http://www.web.archive.org>), o qual indica um volume elevado e persistente de capturas do vídeo entre janeiro de 2016 e setembro de 2022, sugerindo a longevidade de seu impacto simbólico.

3.2 ■ *Etapa 2*

■ *Materialização da pergunta de pesquisa*

Expressão (*query string*) para o motor de busca do *Media Cloud*:

FORMA COMPLETA — ("tartaruga marinha" AND "canudo") OR ("tartaruga marinha" AND "canudos") OR ("tartaruga marinha" AND "lixo plástico") OR ("tartaruga marinha" AND "poluição") OR ("tartaruga marinha" AND "microplásticos") OR ("tartarugas marinhas" AND "canudo") OR ("tartarugas marinhas" AND "canudos") OR ("tartarugas marinhas" AND "lixo plástico") OR ("tartarugas marinhas" AND "poluição") OR

("tartarugas marinhas" AND "microplásticos") OR ("fauna marinha" AND "canudo") OR ("fauna marinha" AND "canudos") OR ("fauna marinha" AND "lixo plástico") OR ("fauna marinha" AND "poluição") OR ("fauna marinha" AND "microplásticos") OR ("vida marinha" AND "canudo") OR ("vida marinha" AND "canudos") OR ("vida marinha" AND "lixo plástico") OR ("vida marinha" AND "poluição") OR ("vida marinha" AND "microplásticos") OR ("animais marinhos" AND "canudo") OR ("animais marinhos" AND "canudos") OR ("animais marinhos" AND "lixo plástico") OR ("animais marinhos" AND "poluição") OR ("animais marinhos" AND "microplásticos")
 FORMA SIMPLIFICADA — ("tartaruga marinha" OR "tartarugas marinhas" OR "fauna marinha" OR "vida marinha" OR "animais marinhos") AND ("canudo" OR "canudos" OR "lixo plástico" OR "poluição" OR "microplásticos")

■ *Bases selecionadas para a coleta*

Coleção #34412257 — “Brazil — National” — 171 fontes —

<http://www.search.mediacloud.org/collections/34412257>

Coleção #38379250 — “Brazil — State & Local” — 1482 fontes —

<http://www.search.mediacloud.org/collections/38379250>

Coleção #8877968 — “Brazil Media” — 1212 fontes —

<http://www.search.mediacloud.org/collections/8877968>

■ *Intervalo de tempo*

01/01/2016–31/12/2022

3.3 ■ *Etapa 3*

A coleta bruta resultou em 3308 notícias. A distribuição das notícias por veículo (Figura 2) revela um ecossistema bem diversificado, que combina grandes veículos generalistas e veículos de nicho especializado. Observa-se, por exemplo, a liderança de veículos voltados à temática ambiental e aos direitos animais, como a ANDA — Agência de Notícias de Direitos Animais, o EcoDebate e o CicloVivo. Veículos de grande alcance, como Globo e UOL, desempenham papel central na difusão do tema para públicos mais amplos, enquanto a presença de veículos regionais, como TNH1 (Alagoas), Bem Paraná (Paraná) e Correio 24 Horas (Bahia), indica uma capilaridade territorial da pauta (região litorânea). Essa diversidade reflete, metodologicamente, o tipo de dado capturado pelo *Media cloud*: Notícias de sites de mídia ao redor da *web*. Cabe ao pesquisador, por exemplo, filtrar apenas veículos que se enquadrem estritamente como espaços jornalísticos (e.g.: Globo) para uma análise mais *stricto sensu*.

A partir desta inspeção inicial (seção 2.2 do *R script*), o TUCANO permite a realização de subanálises focadas em recortes específicos, como a comparação entre veículos especializados e portais de massa.

Após o procedimento de filtragem por idioma e deduplicação (seção 4 e 5 do *R script*), o *corpus* final de análise compreende 2485 notícias únicas. A distribuição cronológica desse *corpus* (Figura 3) evidencia uma intensificação da cobertura a partir de 2018, com pico em 2019, concentrando aproximadamente 29,2% das notícias. Esse padrão corrobora as justificativas apresentadas para a escolha do recorte temporal.

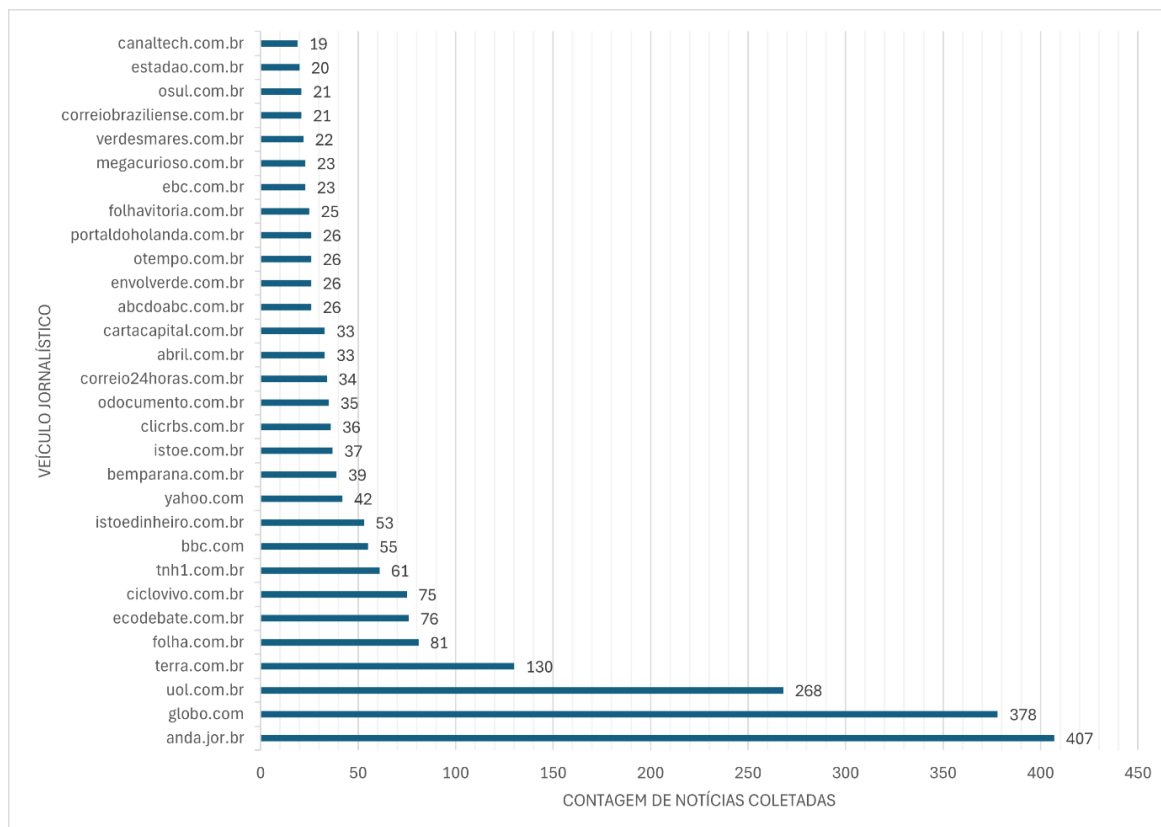


Figura 2. Distribuição da frequência dos 30 veículos com maior volume de notícias no conjunto de dados brutos (N = 3308 notícias). Juntos, os 30 veículos representam aproximadamente 65% das notícias brutas coletadas. Esta análise pode ser realizada após a filtragem de duplicatas na seção 5.3 do *R script*. Figura gerada no Excel com resultados do TUCANO (seção 2.2 do *R script*). Fonte: Elaborado pelos autores, 2026.

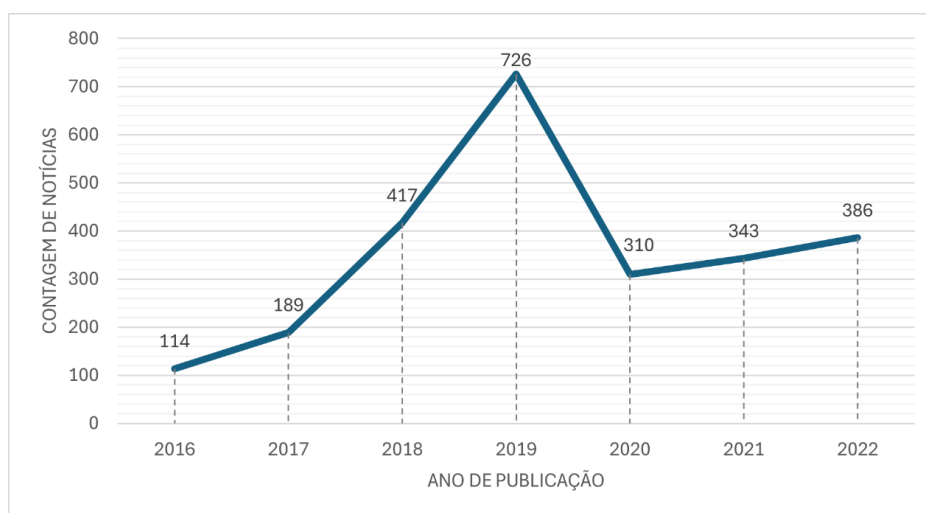


Figura 3. Distribuição cronológica do *corpus* de análise (N = 2485 notícias). Os dados representam o volume de notícias únicas publicadas entre 2016 e 2022 após filtrações operacionais (idioma e deduplicação). Figura gerada no Excel com resultados do TUCANO (seção 5.2 do *R script*). Fonte: Elaborado pelos autores, 2026.

A Etapa 4 constitui o núcleo analítico do TUCANO, pois revela a anatomia textual do *corpus* por meio da tokenização e da representação vetorial dos títulos (seção 6 do *R script*). Os dados foram convertidos em uma Matriz Termo-Documento (seção 6.5 e 6.6 do *R script*) com 4936 termos (*features*) por 2485 notícias (*documents*), estrutura que viabiliza a exploração automatizada do *corpus*.

[illegible]

Embora valores como 2% e 1% aparentam ser baixos, refletem a distribuição estatística típica de análises de linguagem natural. Conforme Jurafsky e Martin [2026], as frequências relativas em sistemas textuais são altamente enviesadas: após a remoção de palavras funcionais (*stopwords*), a massa de ocorrências concentra-se em poucas palavras de conteúdo (*content words*), enquanto a maioria do vocabulário compõe um conjunto de termos raros (*hapax legomena*). A dominância de aproximadamente 2% do termo “plástico”, em um universo de quase 5000 termos, é estruturalmente interessante, ratifica a centralidade temática do *corpus* de análise e valida a potencialidade do TUCANO em isolar núcleos informativamente densos.

Artigo ■ JCOMAL 09(02)(2026)A02 ■ 11

o pesquisador pode realizar a análise de *collocations* (com base nos n-gramas), aplicando filtros de frequência mínima (definidos pelo pesquisador) para identificar sequências de palavras que possuam valor semântico real. O TUCANO oferece liberdade analítica ao permitir o processamento desses n-gramas em dois níveis: os *Collocations* Naturais (seção 10.1 do *R script*), que preservam os *stopwords* para revelar uma estrutura mais “orgânica”, e os *Collocations* Filtrados (seção 10.2 do *R script*), que removem palavras funcionais para isolar de forma mais técnica.

3.5 ■ Etapa 5

Nesta etapa, o protocolo metodológico avança para a análise das relações de coocorrência entre os termos, produzindo cartografias lexicais do *corpus*. Essas redes são construídas a partir da manipulação de dois parâmetros centrais: a janela de contexto (seção 3.2 — *window_fcm* — do *R script*) e a frequência mínima para entrada do termo na rede (seção 3.2 — *min_freq_network* — do *R script*).

A variação desses parâmetros permite ao pesquisador alternar entre análises exploratórias mais densas e visualizações mais sintéticas, capazes de isolar núcleos semânticos estáveis. Nesses modelos de coocorrência, a escolha da janela de contexto influencia a noção de semelhança que o modelo captura [Turney & Pantel, 2010]. Por exemplo, uma janela de contexto igual a 2 foca em relações sintáticas imediatas, enquanto uma janela igual a 10 captura associações mais amplas dentro de um mesmo título. Embora não haja um valor fixo, de modo geral, janelas de contexto locais possibilitam identificar relações lexicais mais finas, enquanto janelas de contexto amplas evidenciam afinidades temáticas e tópicos [Manning & Schütze, 1999]. O filtro de frequência mínima atua como um redutor de ruído, descartando conexões mais periféricas para revelar o “esqueleto” da cartografia.

Essas diferentes configurações, apresentadas na Figura 5, ilustram essa flexibilidade metodológica e demonstram o potencial de redução controlada da dimensionalidade analítica. O pesquisador pode realizar desde uma inspeção exploratória exaustiva (redes mais conectadas, Figuras 5A e 5B) até a identificação de fórmulas mais consolidadas e estáveis (derivação de sub-redes isoladas, Figuras 5C, 5D, 5E e 5F), onde apenas as coocorrências de maior força resistem à filtragem. Em suma, o pesquisador tem a liberdade de gerar a rede de coocorrência quantas vezes desejar, ajustando os parâmetros e observando que diversos agrupamentos de termos vão se evidenciando. Todas essas cartografias podem servir como critério para a subamostragem da etapa 6. Inclusive, é função opcional do TUCANO gerar redes usando termos específicos como âncora da rede (chegar seção 11.OP no *R script*). Por fim, é nessa etapa do TUCANO que o pesquisador também pode gerar arquivos com valores de correlação entre pares de termos (seção 12 do *R script*), que também podem servir de critério de subamostragem com base em valores estatísticos.

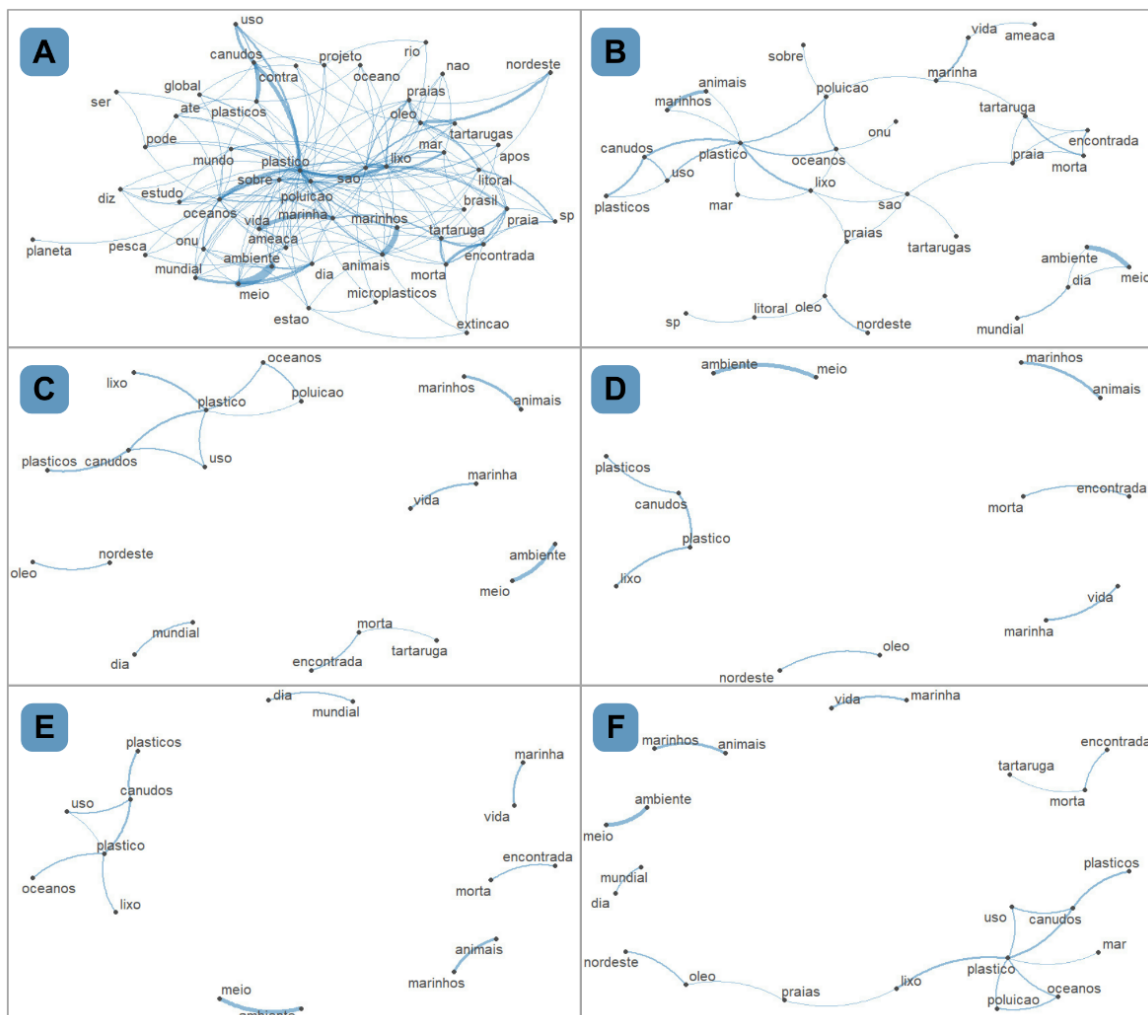


Figura 5. Redes de coocorrência sob diferentes configurações paramétricas (janela de contexto – w ; e a frequência mínima para entrada do termo/token normalizado na rede – mf). As variações demonstram a capacidade de redução de dimensionalidade do protocolo metodológico TUCANO: *A* – $w = 5$, $mf=5$ (Rede de alta densidade); *B* – $w = 5$, $mf=15$; *C* – $w = 5$, $mf=25$; *D* – $w = 5$, $mf=35$ (Isolamento dos núcleos semânticos mais estáveis); *E* – $w = 2$, $mf=25$ (Foco em associações sintáticas curtas e imediatas); *F* – $w = 10$, $mf=25$ (Mapeamento de relações temáticas amplas). Figuras geradas pelo TUCANO no *software R* (seção 11 do *R script*). Fonte: Elaborado pelos autores, 2026.

3.6 ■ Etapa 6

A Etapa 6 sintetiza o principal objetivo do TUCANO ao marcar a transição da exploração técnica de um amplo conjunto de dados para a definição teoricamente orientada de um *corpus* destinado à análise qualitativa. Nesse momento, os resultados obtidos nas Etapas 4 e 5 (como a identificação de termos mais frequentes, bigramas, redes de coocorrência e até estatísticas de correlação) não operam como critérios automáticos de seleção, mas como artefatos analíticos que favorecem a interação contínua do pesquisador com o material empírico. É a partir dessa interação, articulada ao referencial teórico e à lógica específica de cada pesquisa, que o pesquisador define os critérios para a composição do *corpus* qualitativo, mobilizando uma sensibilidade analítica construída no próprio processo de exploração dos dados, e não só pela literatura técnica e experiência profissional. Isso vai de frente com o principal dogma do TUCANO: a ideia de inserir o pesquisador no processo analítico pré-análise qualitativa, através da interação contínua com o material empírico. Acreditamos que isso participa do desenvolvimento de sensibilidade teórica do pesquisador, assim como defendido por Glaser [1978].

Nesta aplicação, adotamos como critério a observação direta das redes de coocorrência identificadas na Etapa 5. A inspeção dessas redes revelou um agrupamento recorrente em torno dos termos “*morta*” e “*encontrada*” (em especial nas Figuras 5C, 5D e 5E). Essa saliência foi interpretada como indicativo de uma oposição semântica fundamental entre vida e morte, que pode ser analisada à luz do nível fundamental (oposições de valores), do Percurso Gerativo de Sentido, da Semiótica Discursiva [Greimas & Courtés, 1979; Fiorin, 1995; Barros, 1988]. Para investigar como esse discurso é construído, a subamostragem temática foi atrelada a esse corpo teórico da semiótica discursiva, focando na narrativa da interrupção da vida (o animal “*morto*” “*encontrado*” como prova do impacto ambiental).

Para operacionalizar essa escolha, os termos “*morta*” e “*encontrada*” foram definidos como as chaves de busca obrigatórias nos títulos das notícias. Na seção 13 do *R script*, é possível escolher entre duas abordagens de subamostragem: Filtragem com lógica AND (seção 13.1) — mantém notícias cujo título contém todas as chaves de busca; e Filtragem com lógica OR (seção 13.2) — mantém notícias cujo título contém pelo menos uma das chaves de busca. Para manter um *corpus* apenas com as notícias cujos títulos contêm os dois termos-chave identificados (“*morta*” e “*encontrada*”), optamos pela filtragem com lógica AND (seção 13.1).

Com isso:

- Critério de recorte: redes de coocorrência da Etapa 5 (Figura 5)
- Termos-chave: “*morta*” e “*encontrada*”
- *Corpus* inicial: 2485 notícias
- Subamostragem final: 43 notícias (Tabela 2)

Este afunilamento evidencia a capacidade do TUCANO de converter padrões matemáticos em objetos de estudo teoricamente fundamentados, funcionando como uma bússola metodológica para a pesquisa qualitativa.

Além de citar este artigo, é importante citar todos os pacotes utilizados pelo TUCANO. São eles: *quanteda* [Benoit et al., 2018], *quanteda.textplots* [Benoit et al., 2018], *quanteda.textstats* [Benoit et al., 2018], *RColorBrewer* [Neuwirth, 2002], *readtext* [Benoit &

Tabela 2. Inventário das notícias selecionadas na subamostragem temática (N = 43). A listagem apresenta os títulos que atenderam aos critérios de filtragem pela lógica AND (“morta” e “encontrada”) a partir do *corpus* inicial (N = 2485). Fonte: Elaborado pelos autores, 2026.

Veículo	Título da notícia (forma exata obtida no Media Cloud)	Ano
afolhatorres.com.br	ICMBio emite comunicado após tartaruga-verde ser encontrada morta na Prainha	2019
anda.jor.br	Cachalotes encontradas mortas na Alemanha tinham estômagos cheios de plástico e peças de carro	2016
anda.jor.br	Em apenas dois dias, duas tartarugas são encontradas mortas em SP	2017
anda.jor.br	Tartaruga é encontrada morta em rede de pesca em SP	2018
anda.jor.br	Tartarugas marinhas são encontradas mortas em praias de Maceió (AL)	2018
anda.jor.br	Tartaruga é encontrada morta em praia de Maceió (AL)	2018
anda.jor.br	Tartaruga é encontrada morta em praia de Fortaleza (CE)	2019
anda.jor.br	Tartaruga é encontrada morta em meio à poluição de canal no RJ	2019
anda.jor.br	Baleia cachalote é encontrada morta com quase 100 kg de plástico no estômago	2019
anda.jor.br	Tartaruga é encontrada morta em praia de Ilhéus (BA)	2019
anda.jor.br	Tartaruga-marinha em extinção é encontrada morta em praia de SP	2022
anda.jor.br	Tartaruga-marinha rara é encontrada morta em Itanhaém (SP)	2021
anda.jor.br	Ameaçada de extinção, tartaruga de 200 kg é encontrada morta em praia de SP	2021
anda.jor.br	Toninha é encontrada morta e instituto suspeita que animal seja vítima da pesca	2021
anda.jor.br	Tartaruga ensanguentada é encontrada morta em praia no litoral da Bahia	2021
anda.jor.br	Ameaçada de extinção, toninha é encontrada morta em praia do litoral de SP	2021
anda.jor.br	Mais de 20 raias ameaçadas de extinção são encontradas mortas em praia	2021
anda.jor.br	Três tartarugas-marinhas de diferentes espécies são encontradas mortas em Praia Grande (SP)	2022
araruna1.com	Baleia é encontrada morta em praia de João Pessoa	2018
atribuna.com.br	Ave rara encontrada morta em Praia Grande ingeriu plástico jogado no mar	2020
bandab.com.br	Tartaruga gigante é encontrada morta por turista em Matinhos; vídeo	2021
bbc.com	A baleia que foi encontrada morta com 40kg de lixo plástico no estômago	2019
diariodolitoral.com.br	Tartaruga caretta caretta é encontrada morta em Peruíbe	2019
folhavoria.com.br	Tartaruga é encontrada morta após ingerir mais de 100 fragmentos de lixo em Vitória	2017
globo.com	Tartaruga é encontrada morta em praia de Pitimbu, na Paraíba	2016
globo.com	Mais uma tartaruga é encontrada morta na Praia de São Conrado	2016
globo.com	Tartarugas são encontradas mortas em Luís Correia e biólogos alertam para lixo jogado na praia	2018
globo.com	Tartaruga é encontrada morta nas areias da praia do Náutico, em Fortaleza	2019
globo.com	Tartaruga é encontrada morta em meio à poluição no Canal da Ogiva, em Cabo Frio, no RJ	2019
globo.com	Tartaruga-verde é encontrada morta em praia de Itajaí	2020
globo.com	Tartaruga é encontrada morta no litoral de Fortaleza	2020
globo.com	Tartaruga com mais de 2 metros e vulnerável à extinção é encontrada morta no litoral de SP	2022
ilheus24h.com.br	SEXTA TARTARUGA MARINHA É ENCONTRADA MORTA NO SUL DA BAHIA	2019
leiaja.com	Cerca de 170 focas são encontradas mortas em três dias	2021
portalmie.com	Tartaruga encontrada morta tinha sacos plásticos no intestino	2020
pradonoticia.com	Prado: Tartaruga marinha é encontrada morta na praia do Novo Prado	2017
r7.com	Plástico no oceano: baleia é encontrada morta com 40kg de sacolas no estômago	2019
terra.com.br	Baleia é encontrada morta com 100 kg de lixo no estômago	2019
tribunahoje.com	Mais de 400 tartarugas são encontradas mortas no litoral alagoano em cinco anos	2016
uol.com.br	Baleia cachalote é encontrada morta com 100 kg de lixo no estômago	2019
verdesmares.com.br	Tartaruga é encontrada morta na praia do Náutico	2019
verdesmares.com.br	Baleia; encontrada morta com 100 kg de lixo no est; mago	2019
yahoo.com	Baleia é encontrada morta com 29 quilos de plástico no estômago na Espanha	2018

Obeng, 2017], *stringi* [Gagolewski, 2022], *stringr* [Wickham, 2009], *tidyverse* [Wickham et al., 2019] e *tm* [Feinerer et al., 2008; Feinerer & Hornik, 2007]. É possível obter todas as referências automáticas na seção 14 do R *script* do TUCANO.

4 - Limitações e considerações finais

Este artigo apresentou o TUCANO — Técnicas Unificadas de Categorização e Análise de Notícias Online — como uma resposta metodológica a uma lacuna central das Ciências Sociais aplicadas na era digital: a dificuldade operacional de construir *corpora* textuais empíricos e coerentes a partir de grandes volumes de dados midiáticos. O protocolo metodológico atua precisamente nesse ponto, oferecendo um dispositivo de escalonamento analítico que permite ao pesquisador transitar de um campo empírico amplo para recortes progressivamente mais precisos e teoricamente orientados.

A aplicação do TUCANO à cobertura midiática sobre canudos plásticos e vida marinha demonstrou seu potencial como ferramenta híbrida, capaz de articular técnicas do Processamento de Linguagem Natural (NLP) com uma lógica interpretativa oriunda das Ciências Sociais. Ao final do fluxo metodológico, a definição do *corpus* deixa de ser uma escolha guiada apenas por conveniência ou intuição e passa a constituir uma decisão informada por uma cartografia prévia do território empírico analisado.

Entretanto, é importante destacar algumas limitações epistemológicas do TUCANO. Primeiramente, o *corpus* analisado é capturado no *Media Cloud*, base de dados ampla e operacionalmente estratégica, mas que não representa a totalidade do universo midiático relacionado ao tema investigado. Os materiais recuperados constituem um recorte sobre o assunto, e não uma representação exaustiva do debate público.

O TUCANO analisa títulos jornalísticos, devido à limitação de direitos autorais de acesso a matérias completas. Embora títulos concentrem estratégias de enquadramento, síntese e direcionamento de leitura [Van Dijk, 1988], compreensões mais profundas acerca dos regimes de sentido, disputas discursivas e frames exigem necessariamente a análise integral das notícias. Os resultados produzidos pelo TUCANO não devem ser interpretados como conclusões qualitativas automatizadas sobre o discurso, mas apenas como uma etapa exploratória de organização e subamostragem do *corpus*. A ampliação do seu uso em textos integrais, inclusive com ferramentas computacionais mais adequadas para textos com maior complexidade sintática e semântica do que os títulos, é um dos caminhos de futuro aperfeiçoamento do TUCANO.

Esta ponte permite situar o TUCANO frente a desenvolvimentos recentes da área do Processamento de Linguagem Natural (NLP). Embora o protocolo utilize técnicas difundidas na literatura do NLP, parte de suas operações pode parecer, em primeira instância, menos sofisticada quando comparada a abordagens contemporâneas baseadas em métodos automatizados capazes de modelar relações semânticas contextuais e agrupamentos temáticos em *corpora* complexos, como os baseados em *embeddings* semânticos. Contudo, o objetivo do TUCANO distancia-se epistemologicamente de propostas orientadas à automatização interpretativa do texto. O TUCANO não busca substituir a leitura analítica do pesquisador por inferências autônomas, mas opera como um dispositivo de exploração do *corpus*, preservando a centralidade da interpretação humana na análise discursiva. As ferramentas computacionais empregadas atuam menos como mecanismos de classificação definitiva e mais como operadores de aproximação, organização e navegação sobre grandes volumes de textos.

Quanto à busca e à coleta do *corpus*, nas etapas 1 e 2, *query strings* mais amplas tendem a ampliar a cobertura do *corpus* recuperado, embora possam reduzir a precisão em relação a recortes mais específicos. As subamostragens produzidas pelo TUCANO não devem ser compreendidas como representações exaustivas do universo empírico analisado, mas como estratégias de dimensionamento exploratório do material coletado. Por exemplo, na Tabela 2, algumas notícias não tratam diretamente da relação entre canudos plásticos, poluição marinha e tartarugas — algumas abordam baleias, aves ou até mesmo outros contextos. Cabe ao pesquisador decidir se trabalhará com a totalidade das notícias subamostradas ou se realizará amostragem por intencionalidade orientada pelos objetivos da sua pesquisa.

O TUCANO contribui para que a interpretação humana permaneça no centro da investigação científica, mesmo em contextos de sobrecarga informacional, reafirmando o papel crítico das

Ciências Sociais frente aos desafios da era digital. Além disso, ao adotar uma arquitetura técnica leve e acessível, o TUCANO contribui para a popularização do uso de ferramentas computacionais em ensino e pesquisa sobre Mídias digitais, Jornalismo e Comunicação, ampliando o repertório metodológico disponível aos profissionais da área.

5 - Disponibilidade de dados

Os dados brutos utilizados na aplicação modelo do TUCANO, bem como o R *script*, estão disponíveis em repositório público no GitHub (<http://www.github.com/victorfelipe62vf/TUCANO>). Como material complementar, está disponível um chat baseado em Inteligência Artificial, NotebookLM (Google), com o objetivo de auxiliar na compreensão e operacionalização das etapas metodológicas. O acesso ao chat pode ser realizado em: <https://doi.org/10.5281/zenodo.18332942>.

Agradecimentos

Os autores gostariam de agradecer às seguintes fontes de financiamento: CAPES (0622/2020/88881.506612/2020-01); Pró-Reitoria de Pesquisa da Universidade Federal de Minas Gerais; Conselho Nacional de Desenvolvimento Científico e Tecnológico (303688/2025-4); Fundação de Amparo à Pesquisa do Estado de Minas Gerais (APQ-04900-22 091/2022 — PDPG III — Parcerias Estratégicas — Edital CAPES 38/2022; APQ-01523-24); Instituto Nacional de Ciência e Tecnologia em Dengue e Interação Microrganismo-Hospedeiro (CNPq 408527/2024-2); Instituto Nacional de Ciência e Tecnologia Patógenos de Doenças Humanas Negligenciadas (INCT-PDHN — CNPq, 408678/2024-0); Rede de Pesquisa em Imunobiológicos e Biofármacos para Terapias Avançadas e Inovadoras (ImunoBioFar), com apoio da FAPEMIG (números de concessão RED-00202-22, APQ-29568-1, APQ-02618-23). FMS é bolsista do CNPq. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — Brasil (CAPES) — Código de Financiamento 001. Agradecemos Verônica, Mauricio, Adlane e Vanessa pelas fecundas sugestões e discussões sobre o texto. Utilizou-se do ChatGPT (OpenAI) como suporte para compreender erros no R *script* durante sua elaboração.

Referências

- Assembleia Legislativa do Estado de São Paulo. (2019). *Lei nº 17.123, de 25 de junho de 2019. Proíbe o fornecimento de canudos de material plástico aos clientes de hotéis, restaurantes, bares, padarias e outros estabelecimentos, e dá providências correlatas.* <https://legislacao.prefeitura.sp.gov.br/leis/lei-17123-de-25-de-junho-de-2019>
- Assembleia Legislativa do Estado do Rio de Janeiro. (2018). *Lei nº 8.006, de 25 de junho de 2018. Dispõe sobre a proibição de canudos plásticos e dá outras providências. Assembleia Legislativa do Estado do Rio de Janeiro.* <https://leisestaduais.com.br/rj/lei-ordinaria-n-8006-2018-rio-de-janeiro-modifica-a-lei-n-5C%C2%5C%BA-5502-de-15-de-julho-de-2009-que-dispoe-sobre-a-substituicao-e-recolhimento-de-sacolas-plasticas-em-estabelecimentos-comerciais-localizados-no-estado-do-rio-de-janeiro-como-forma-de-coloca-las-a-disposicao-do-ciclo-de-reciclagem-e-protecao-ao-meio-ambiente-fluminense>
- Barros, D. L. P. d. (1988). *Teoria do discurso: Fundamentos semióticos*. Atual.

- Bauer, M. W., & Gaskell, G. (2017). *Pesquisa qualitativa com texto, imagem e som: Um manual prático*. Editora Vozes.
- Benoit, K., & Obeng, A. (2017). *readtext: Import and Handling for Plain and Formatted Text Files (Versão 0.92.1)* [Pacote de R]. <https://doi.org/10.32614/cran.package.readtext>
- Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: an R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30), 774. <https://doi.org/10.21105/joss.00774>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022. <https://doi.org/10.5555/944919.944937>
- Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662–679. <https://doi.org/10.1080/1369118x.2012.678878>
- Castelfranchi, Y. (2017). Computer-aided text analysis: an open-aided laboratory for social sciences. *JCOM*, 16(02), C04. <https://doi.org/10.22323/2.16020304>
- Castells, M. (2009). *The Rise of the Network Society* (2ª ed.). Wiley. <https://doi.org/10.1002/9781444319514>
- Feinerer, I., & Hornik, K. (2007). tm: Text Mining Package (Versão 0.7-16) [Pacote de R]. CRAN: Contributed Packages. <https://doi.org/10.32614/cran.package.tm>
- Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, 25(5), 1–54. <https://doi.org/10.18637/jss.v025.i05>
- Figgenger, C. (2015). *Sea turtle with straw up its nostril – “NO” to single-use plastic* [Video]. <https://www.youtube.com/watch?v=4wH878t78bw>
- Fiorin, J. L. (1995). A noção de texto na semiótica. *Organon*, 9.
- Gagolewski, M. (2022). stringi: Fast and Portable Character String Processing in R. *Journal of Statistical Software*, 103(2), 1–59. <https://doi.org/10.18637/jss.v103.i02>
- Gil, A. C. (2024). *Métodos e técnicas de pesquisa social* (7ª ed.). Editora Atlas.
- Glaser, B. G. (1978). *Theoretical sensitivity: Advances in the methodology of grounded theory*. Sociology Press.
- Gould, L., & Rocha, V. (2021). O uso social dos afetos: um olhar sobre o jornalismo sensível na era da desinformação. *Comunicologia - Revista de Comunicação da Universidade Católica de Brasília*, 14(1), 102. <https://doi.org/10.31501/comunicologia.v14i1.12814>
- Greimas, A. J., & Courtés, J. (1979). *Dicionário de semiótica* (A. D. Lima et al., Trans.). Editora Cultrix.
- Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3), 267–297. <https://doi.org/10.1093/pan/mps028>
- Hallin, D. C., & Mancini, P. (2004). *Comparing Media Systems: Three Models of Media and Politics*. Cambridge University Press. <https://doi.org/10.1017/cbo9780511790867>
- Hissa, D. L. A. (2024). A leitura plataformizada na mídia digital: da sobrecarga informacional à perda do pensamento crítico. *Revista da ABRALIN*, 304–331. <https://doi.org/10.25189/rabralin.v23i2.2208>
- Jurafsky, D., & Martin, J. H. (2026). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models* (3ª ed.). Stanford University. <https://web.stanford.edu/~jurafsky/slp3>
- Leal, B. S., Antunes, E., & Vaz, P. B. (2014). *Para entender o jornalismo*. Autêntica.

- Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. <https://doi.org/10.1017/cbo9780511809071>
- Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. The MIT Press. <https://mitpress.mit.edu/9780262303798/foundations-of-statistical-natural-language-processing/>
- Medina, C. (2002). *Ciência e jornalismo: Da herança positivista ao diálogo dos afetos*. ECA/USP.
- Moreira, D., Pires, A., & Medeiros, M. d. A. (2022). Do ‘texto como texto’ ao ‘texto como dado’: o potencial das pesquisas em Relações Internacionais. *Revista de Sociologia e Política*, 30, e005. <https://doi.org/10.1590/1678-98732230e005>
- Neuwirth, E. (2002). RColorBrewer: ColorBrewer Palettes (Versão 1.1-3) [Pacote de R]. CRAN: Contributed Packages. <https://doi.org/10.32614/cran.package.rcolorbrewer>
- Neves, L. F. F., & Massarani, L. (2022). Politics, economy and society in the coverage of COVID-19 by elite newspapers in US, UK, China and Brazil: a text mining approach. *JCOM*, 21(07), A04. <https://doi.org/10.22323/2.21070204>
- Nunes, E. D. (2007). O desafio do conhecimento: pesquisa qualitativa em saúde. *Ciência & Saúde Coletiva*, 12(4), 1087–1088. <https://doi.org/10.1590/s1413-81232007000400030>
- Pena, F. (2021). *Teoria do jornalismo* (3ª ed.). Editora Contexto.
- Posit Team. (2025). *RStudio: Integrated development environment for R* [Software de computador]. <http://www.posit.co/>
- R Core Team. (2025). *R: a language and environment for statistical computing* [Software de computador]. <https://www.R-project.org/>
- Roberts, H., Bhargava, R., Valiukas, L., Jen, D., Malik, M. M., Bishop, C. S., Ndulue, E. B., Dave, A., Clark, J., Etling, B., Faris, R., Shah, A., Rubinovitz, J., Hope, A., D’Ignazio, C., Bermejo, F., Benkler, Y., & Zuckerman, E. (2021). Media Cloud: Massive Open Source Collection of Global News on the Open Web. *Proceedings of the International AAAI Conference on Web and Social Media*, 15, 1034–1045. <https://doi.org/10.1609/icwsm.v15i1.18127>
- Sampaio, R. C., Sanchez, C. S., Marioto, D. J. F., Araujo, B. C. d. S., Herédia, L. H. O., Paz, F. S., Tigrinho, C. S., & Souza, J. R. d. (2022). Muita Bardin, pouca qualidade: uma avaliação sobre as análises de conteúdo qualitativas no Brasil. *Revista Pesquisa Qualitativa*, 10(25), 464–494. <https://doi.org/10.33361/rpq.2022.v.10.n.25.547>
- Santos De Mattos, F. (2023). Plataformização das Notícias: análise dos níveis de presença plataformizada dos media jornalísticos portugueses. *Contratexto*, (039), 23–49. <https://doi.org/10.26439/contratexto2023.n39.6011>
- Schudson, M. (1981). *Discovering the news: A social history of American newspapers*. Basic Books.
- Senado Federal do Brasil. (2018). *Projeto de Lei do Senado nº 92, de 2018. Altera a Lei nº 12.305, de 2 de agosto de 2010, para proibir a fabricação, a importação, a distribuição e a comercialização de canudos plásticos e outras providências*. <https://www25.senado.leg.br/web/atividade/materias/-/materia/132457>
- Souza, G. L. d., & Coutinho, I. M. d. S. (2025). Amenazas algorítmicas al periodismo: Las lógicas de las plataformas digitales y la inteligencia artificial como (no) referencias para la producción y circulación de noticias. *Razón y Palabra*, 29, 5–15. <https://doi.org/10.26807/rp.v29i122.2211>
- Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, 141–188. <https://doi.org/10.48550/arXiv.1003.1141>
- van Dijck, J. (2013). *The Culture of Connectivity: A Critical History of Social Media*. Oxford University Press, New York. <https://doi.org/10.1093/acprof:oso/9780199970773.001.0001>
- Van Dijck, T. A. (1988). *News as discourse*. Lawrence Erlbaum Associates.

- van Loon, A. (2022). Three families of automated text analysis. *Social Science Research*, 108, 102798. <https://doi.org/10.1016/j.ssresearch.2022.102798>
- Walsh-Childers, K., & McKinnon, M. (2024). Why Science and Health Journalism Matters. Em *Palgrave Handbook of Science and Health Journalism* (pp. 1–13). Springer International Publishing. https://doi.org/10.1007/978-3-031-49084-2_1
- Wickham, H. (2009). stringr: Simple, Consistent Wrappers for Common String Operations (Versão 1.6.0) [Pacote de R]. CRAN: Contributed Packages. <https://doi.org/10.32614/cran.package.stringr>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., ... Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>

Sobre os autores

Victor Felipe Mendanha de Lima. Sou biólogo, mestre em Botânica e estou finalizando meu doutorado em Genética pela Universidade Federal de Minas Gerais (UFMG), atuando na interface entre Ciência, Linguística e Comunicação pública do conhecimento científico. Minha atuação combina formação científica avançada com estratégias contemporâneas de comunicação em mídias digitais, transformando conhecimento altamente técnico em conteúdos criativos, envolventes e estrategicamente posicionados para diferentes públicos. Trabalho com produção e estratégia de conteúdo científico em mídias digitais, integrando storytelling, escrita criativa, análise de discurso e marketing de conteúdo para comunicar temas complexos em Genética, Biologia Molecular, Biotecnologia, Saúde e Ciência contemporânea.



victorfelipe62.vf@gmail.com

Frederico Marianetti Soriani. Professor Doutor Associado do Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas — UFMG. Farmacêutico, coordenador do Laboratório de Genética Funcional, pesquisador na área de imunologia com atuação em Divulgação Científica.



fredsori@gmail.com

Como citar

Mendanha de Lima, V. e Soriani, F. (2026). 'Como estudar mídias digitais em tempos de dilúvio informacional? O protocolo metodológico TUCANO aplicado à pesquisa documental em notícias online'. *JCOMAL* 09(02), A02. <https://doi.org/10.22323/392520260724172353>.



© O(s) autor(es).

Esta publicação é disponibilizada nos termos da licença Creative Commons [Atribuição 4.0](https://creativecommons.org/licenses/by/4.0/). Todos os direitos de Mineração de Texto e Dados, treinamento de IA e tecnologias similares para fins comerciais são reservados.

ISSN 2611-9986. Publicado pela SISSA Medialab. jcomal.sissa.it