Anais do Congresso Brasileiro de Iniciação Científica
(https://periodicos.unisanta.br/COB)

ISSN 2317-1316

PROCESSAMENTO DE LINGUAGEM NATURAL E ANÁLISE DISCURSIVA: UMA ABORDAGEM PARA A ANÁLISE DE TEXTOS JORNALÍSTICOS

NATURAL LANGUAGE PROCESSING AND DISCOURSE ANALYSIS: AN APPROACH TO THE ANALYSIS OF JOURNALISTIC TEXTS

Guilherme Onorio Pereira da Silva, Denise Durante

Faculdade de Tecnologia da Baixada Santista Rubens Lara, Curso de Ciência de Dados
guilherme.silva22@aluno.cps.sp.gov.br; denise.durante@cps.sp.gov.br

RESUMO – O Processamento de Linguagem Natural (PLN) tem se consolidado como ferramenta para a análise de grandes volumes de textos, ampliando as possibilidades de investigação dos discursos presentes em veículos digitais de comunicação. Ao relacionar a Linguística de Corpus à Análise do Discurso, o PLN pode contribuir para a identificação de estratégias discursivas, conteúdos ideológicos e relações de poder subjacentes aos textos. Nesse sentido, este projeto tem como objetivo descrever e analisar as potencialidades do PLN para contribuir com a Análise do Discurso, ampliando a coleta de dados e aprofundando a análise dos sentidos construídos em textos jornalísticos e de opinião. A pesquisa fundamenta-se na vertente francesa da Análise do Discurso, especialmente nos estudos de Pêcheux (1990), Maingueneau (2015; 2018) e Charaudeau (2004). Para isso, foram construídos bancos de dados textuais a partir de acervos jornalísticos e aplicados algoritmos de bibliotecas da linguagem Python. Os resultados permitiram identificar, descrever e analisar contribuições do PLN para os analistas do discurso no contexto da comunicação digital.

Palavras-chave: Análise do Discurso; Ciência de Dados; Processamento de Linguagem Natural; Comunicação de Massa; Jornalismo.

ABSTRACT – Natural Language Processing (NLP) has become an increasingly relevant tool for analyzing large volumes of text, expanding the possibilities for investigating discourses in digital media. By integrating Corpus Linguistics and Discourse Analysis, NLP can contribute to identifying discursive strategies, ideological content, and power relations underlying textual production. In this context, this study aims to describe and analyze the potential of NLP to support Discourse Analysis by expanding data collection and deepening the investigation of meaning-making processes in journalistic and opinion texts. The research is grounded in the French tradition of Discourse Analysis, particularly the studies of Pêcheux (1990), Maingueneau (2015, 2018), and Charaudeau (2004). To this end, textual datasets were compiled from journalistic archives, and algorithms

ISSN 2317-1316

implemented through Python libraries were applied. The results enabled the identification, description, and analysis of the contributions of NLP to discourse analysts in the context of digital communication, highlighting its potential to expand empirical investigation and support the analysis of discursive phenomena in large textual corpora.

Keywords: Discourse Analysis; Data Science; Natural Language Processing; Mass Communication; Journalism.

1 INTRODUÇÃO

A Inteligência Artificial (IA) tem se tornado uma ferramenta prática na vida dos indivíduos. Sendo assim, diversas ferramentas vêm sendo desenvolvidas para contribuir com atividades cotidianas nas sociedades em que a Tecnologia de Informação (TI) exerce papel fundamental, sendo o impulsionamento de Processamento de Linguagem Natural (PLN) amplamente utilizado para a realização de tarefas que requerem dados extraídos de textos falados ou escritos (JURAFSKY, MARTIN, 2024).

Essa ferramenta, enquanto área da tecnologia dedicada ao desenvolvimento de métodos capazes de processar linguagem humana, amplia suas possibilidades de aplicação ao estabelecer interfaces com outros campos do conhecimento, como a Educação (FELIPPO, DIAS-DA-SILVA, 2009).

Em meio a esse cenário, alguns questionamentos surgem quando a aplicação da IA ocorre dentro de um ambiente comunicacional. A especificidade da linguagem humana, explorada pelos estudos da Linguística e da Análise do Discurso, pode ser pensada no contexto da IA generativa de texto. Desse modo, devem-se considerar aspectos da fonologia, da morfologia, da sintaxe e da semântica de sentenças. Isso inclui, por exemplo, a reflexão sobre ambiguidades, gírias comuns e de grupos, bem como de usos da língua marcados por variações diacrônicas, diatópicas, diastráticas e diafásicas. Podem-se considerar necessários, portanto, critérios para a identificação de modelos adequados para a avaliação sobre sentidos implícitos nos textos gerados pela IA.

Michel Pêcheux (1990), no contexto da Análise do Discurso de linha francesa, propõe que uma leitura que leva a "uma forma de conhecimento em que se faz necessária a oposição, a contradição entre teoria e sua prática de análise". O autor aponta para a ausência de

ISSN 2317-1316

neutralidade da língua e a influência de poder que alguns setores podem ter na sociedade por meio dos usos da linguagem.

A aplicação de técnicas de PLN para os estudos da Análise do Discurso pode oferecer a oportunidade para explorar o tema sobre como as tecnologias de processamento de linguagem natural podem ser usadas para decifrar, mapear e analisar as estratégias discursivas presentes (MELO, 2009).

Diante desse contexto, este estudo tem como objetivo geral compreender métodos de Processamento de Linguagem Natural (PLN) em gêneros textuais como artigos de opinião e notícias jornalísticas, sob a perspectiva da Análise do Discurso, desenvolvida por Michel Pêcheux, Patrick Charaudeau e Dominique Maingueneau. Para isso, serão exploradas técnicas de modelagem de tópicos (*Topic Modeling*) e análise de dependências sintáticas, considerando suas possibilidades e limitações na interpretação dos sentidos produzidos nos textos.

2 MATERIAIS E MÉTODOS

Este trabalho faz parte de uma pesquisa de Iniciação Científica na área de Ciência de Dados. Adota-se na pesquisa uma abordagem metodológica híbrida, integrando procedimentos quantitativos e qualitativos com a articulação entre técnicas de Processamento de Linguagem Natural (PLN) e os pressupostos da Análise do Discurso (AD). Essa combinação metodológica parte do entendimento de que métodos computacionais possibilitam a identificação de padrões recorrentes em grandes volumes de dados textuais (BERBER SARDINHA, 2004), enquanto a Análise do Discurso oferece base conceitual para a interpretação sobre os efeitos de sentido produzidos pelas escolhas linguísticas e pelas condições de produção dos enunciados (MELO, 2009). Dessa forma, busca-se compreender não apenas quais temas são recorrentes nos textos jornalísticos, mas também como esses temas são organizados linguisticamente e construídos discursivamente.

O *corpus* da pesquisa é composto por quarenta notícias e editoriais publicados em veículos jornalísticos de circulação nacional e regional. A aplicação do algoritmo foi feita em amostras de cinco textos. A escolha desses gêneros textuais fundamenta-se na relevância que desempenham na circulação de discursos sociais e políticos, uma vez que tanto a notícia quanto o editorial apresentam diferentes graus de posicionamento enunciativo. Dentro do *ethos*

ISSN 2317-1316

jornalístico, a notícia tende a produzir um efeito de objetividade, o editorial explicita a posição institucional do veículo de comunicação (TRAQUINA, 2005).

Na etapa quantitativa, foi empregada a técnica de *Topic Modeling*, responsável pela identificação automática de temas recorrentes presentes no *corpus* (GROOTENDORST, 2022). Inicialmente, foram realizados procedimentos de pré-processamento textual, incluindo tokenização, lematização e remoção de elementos considerados pouco informativos. Em seguida, aplicaram-se modelos probabilísticos e semânticos para identificar agrupamentos lexicais capazes de representar os principais assuntos abordados em cada documento. A modelagem de tópicos permite reconhecer regularidades discursivas sem depender exclusivamente de categorias previamente definidas, favorecendo análises exploratórias em grandes conjuntos documentais.

Para a implementação computacional, foram utilizadas bibliotecas em linguagem de programação *Python* com utilização na área de Processamento de Linguagem Natural (MOREIRA FILHO, 2021). A biblioteca *spaCy* foi empregada nas etapas de tokenização, lematização, reconhecimento de entidades nomeadas, análise sintática e geração das árvores de dependências. A biblioteca *Gensim* foi utilizada para a construção dos modelos probabilísticos de tópicos por meio do algoritmo *Latent Dirichlet Allocation* (LDA), o que possibilitou identificar distribuições temáticas no *corpus* (BLEI, NG, JORDAN, 2003). Paralelamente, utilizou-se a biblioteca *BERTopic*, baseada em representações vetoriais (*embeddings*) e técnicas de agrupamento semântico, permitindo capturar relações de significado que extrapolam a simples frequência lexical. Foram utilizadas ferramentas de Inteligência Artificial para a revisão e correção do algoritmo empregado para a análise computacional.

3 RESULTADOS E DISCUSSÃO

A pesquisa considera as bases teóricas dos autores selecionados para o desenvolvimento do estudo, além da evolução do escopo da investigação e da definição do *corpus* utilizado para a aplicação dos métodos de Processamento de Linguagem Natural (PLN), fundamentados nos principais conceitos da Análise do Discurso de linha francesa. A bibliografia consultada foi essencial para a construção da metodologia, visto que permitiu identificar padrões descritos principalmente por Michel Pêcheux e relacioná-los à análise empírica realizada com o auxílio

ISSN 2317-1316

de algoritmos de PLN. Esses algoritmos possibilitaram a identificação quantitativa de palavras, bem como a análise de suas funções morfológicas e sintáticas, fornecendo informações relevantes para a interpretação dos textos.

O desenvolvimento do modelo de PLN baseia-se na teoria da Análise do Discurso, tendo como ponto de partida a construção de um banco de dados textual composto por notícias e editoriais publicados nos jornais *Folha de S. Paulo*, *O Estado de S. Paulo*, *O Globo* e *A Tribuna*. Foram selecionados quarenta textos relacionados às temáticas ESG (*Environmental*, *Social and Governance*) e à editoria de política.

Observou-se que a aplicação dos conceitos da Análise do Discurso em modelos de Processamento de Linguagem Natural apresenta desafios significativos. A principal dificuldade consiste em traduzir conceitos teóricos complexos, como hegemonia, ideologia e influência discursiva, em parâmetros computacionais capazes de serem processados por algoritmos. Além disso, a identificação de elementos implícitos, pressupostos e nuances discursivas representa um obstáculo adicional, uma vez que esses aspectos frequentemente extrapolam a análise estatística da linguagem. Por esse motivo, a metodologia foi organizada em etapas, separando inicialmente a distribuição linguística produzida pelo algoritmo e, posteriormente, a interpretação individual fundamentada nos pressupostos da Análise do Discurso.

Em um banco de dados composto por aproximadamente 40 notícias e editoriais, foram separadas cinco amostras, sendo três, as que obtiveram resultados mais próximos por meio da combinação entre *Topic Modeling* e Análise de Dependências, permitindo integrar a capacidade computacional de identificação de padrões linguísticos à análise qualitativa realizada pelo pesquisador. Enquanto a modelagem de tópicos evidencia os principais campos semânticos presentes nos documentos, a análise sintática revela como esses temas são organizados pelas relações gramaticais estabelecidas nas orações. A interpretação baseada na Análise do Discurso foi realizada em uma etapa posterior, funcionando como principal parâmetro qualitativo de avaliação.

Na notícia “Atrito entre Kassab e Tarcísio esvazia apoio do PSD no interior paulista”, publicada pela *Folha de S. Paulo* (RIBEIRO, 2026), a modelagem de tópicos identificou como termos centrais expressões relacionadas ao contexto eleitoral, como “primeiro turno”, “segundo turno”, “reeleição”, “Senado”, “pré-campanha” e “palanques”. Entre os termos recorrentes,

ISSN 2317-1316

destacaram-se “atrito”, “apoio”, “vice”, “disputa”, “aliados”, “campanha” e “enfraquecê-lo”, evidenciando um campo lexical associado à competição política e às negociações partidárias. Os termos complementares, como “governador”, “Palácio dos Bandeirantes”, “Assembleia Legislativa” e “prefeitos”, reforçam a centralidade das instituições políticas e dos agentes envolvidos nas disputas eleitorais.

A Análise de Dependências aplicada à frase “Prefeitos e deputados do PSD avaliam não trabalhar pela campanha de Tarcísio” demonstrou que o sujeito da oração é formado pelos agentes políticos responsáveis pela ação verbal “avaliam”, enquanto a estrutura infinitiva “não trabalhar” representa o conteúdo da decisão. A presença explícita da negação desloca o foco do enunciado para a ausência de apoio político, produzindo um efeito discursivo de enfraquecimento da candidatura. Esse resultado pode evidenciar como a estrutura sintática pode complementar a análise temática ao revelar relações que não seriam identificadas apenas pela consideração da frequência lexical.

Na notícia publicada pelo jornal *A Tribuna* (2026) com o título de “Praia de Santos tem diminuição na faixa de areia mais evidente nas últimas semanas”, o *Topic Modeling* identificou tópicos relacionados às questões ambientais e costeiras. Entre os principais termos encontrados estão “nível do mar”, “ressacas”, “erosão”, “correntes”, “sedimentos”, “ondas” e “faixa de areia”. Também foram identificados termos recorrentes relacionados à urbanização, como “concreto”, “manguezais”, “restinga” e “estruturas de contenção”. Ao analisar a sentença “O mar está clamando de volta parte daquela região”, a Análise de Dependências evidenciou uma construção metafórica ao identificar “mar” como sujeito da oração, atribuindo características de agente a um fenômeno natural e intensificando discursivamente a percepção da erosão costeira.

No editorial de *Notas & Informações* (2026) intitulado “O Brasil escolheu o atraso”, publicado pelo jornal *O Estado de S. Paulo*, a modelagem de tópicos identificou predominância de termos associados ao cenário político e econômico, como “eleição”, “populismo”, “Congresso”, “governo”, “tributário”, “investimentos”, “infraestrutura”, “produtividade”, “PIB”, “desenvolvimento” e “crise”. Na frase “O Brasil cresce em movimentos espasmódicos”, a Análise de Dependências demonstrou que o país ocupa a posição de sujeito da oração,

ISSN 2317-1316

enquanto a expressão “em movimentos espasmódicos” caracteriza metaforicamente o crescimento econômico como irregular e descontínuo.

Também foi observado que, caso existam erros na estruturação do algoritmo ou da captação do texto no banco de dados, podem existir retornos que não se adequam como visto em verbos sendo indicados como adjetivos e substantivos ou conectar sentenças em n-gramas e a remoção de *stopwords* para a saída adequada dentro do contexto. Assim sendo, para retratar "nível do mar", "faixa de areia", "Assembleia Legislativa", a necessidade de analisar os resultados, adequá-los aos contextos que estavam a partir da leitura do material e repassá-los a utilização de Análise de Discurso.

4 CONCLUSÃO

Os resultados obtidos demonstram que a integração entre técnicas de Processamento de Linguagem Natural (PLN) e os pressupostos da Análise do Discurso constitui uma abordagem viável para a investigação de textos jornalísticos e de opinião. A aplicação conjunta de *Topic Modeling* e Análise de Dependências possibilitou identificar recorrências temáticas, padrões lexicais e relações sintáticas, fornecendo subsídios para a análise dos sentidos construídos nos textos. Nesse processo, o PLN mostrou-se relevante como instrumento de apoio à exploração sistemática de grandes volumes de dados presentes em textos, o que permite localizar regularidades que podem orientar a interpretação discursiva. Entretanto, os resultados também evidenciaram que a identificação computacional de padrões não é suficiente para compreender integralmente aspectos relacionados às condições de produção, à ideologia, à memória discursiva e aos sentidos implícitos. Dessa forma, a interpretação qualitativa permanece fundamental para a compreensão dos fenômenos discursivos. Conclui-se que a articulação entre métodos computacionais e os fundamentos da Análise do Discurso amplia as possibilidades de investigação de *corpora* jornalísticos, conciliando a capacidade de processamento do PLN com a interpretação crítica necessária à análise dos discursos.

5 DECLARAÇÃO DE USO DE IA

Durante a elaboração deste trabalho, utilizou-se a ferramenta *Claude Sonnet* 5.5 para auxiliar a correção e a revisão do algoritmo empregado na análise sintática e no processamento

ISSN 2317-1316

estatístico dos dados. O conteúdo gerado foi revisado, editado e validado pelos autores, que assumem total responsabilidade pelo texto final.

6 REFERÊNCIAS

ATRIBUNA.COM.BR. Praia de Santos tem diminuição na faixa de areia mais evidente nas últimas semanas; avanço do mar não é única causa. A Tribuna, Santos, 21 abr. 2026. Disponível em: https://www.atribuna.com.br/cidades/santos/praia-de-santos-tem-diminuic-o-na-faixa-de-areia-mais-evidente-nas-ultimas-semanas-avanco-do-mar-n-o-e-unica-causa-1.510724. Acesso em: 13 out. 2026.

BERBER SARDINHA, T. Linguística de *Corpus*. Barueri, SP. Manole, 2004.

BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent Dirichlet Allocation. Journal of Machine Learning Research, Cambridge, v. 3, p. 993–1022, 2003. Disponível em: https://dl.acm.org/doi/10.5555/944919.944937. Acesso em: 30 jun. 2026.

FELIPPO, A. Di, DIAS-DA-SILVA, B. C. O processamento automático de línguas naturais enquanto engenharia do conhecimento linguístico. In: Caleidoscópio, v.7, n.3, p.183-191. São Leopoldo, RS. Unisinos, 2009.

GROOTENDORST, Maarten. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint, arXiv:2203.05794, Computation and Language, 2022. Disponível em: https://arxiv.org/abs/2203.05794. Acesso em: 30 jun. 2026.

JURAFSKY, Daniel; MARTIN, James H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models. 3. ed. Draft, 2024. Disponível em: https://web.stanford.edu/~jurafsky/slp3/ed3bookaug20_2024.pdf. Acesso em: 30 jun. 2026.

MELO, I. Análise de Discurso e Análise Crítica de Discurso: Desdobramentos e Intersecções. Revista Eletrônica de Divulgação Científica em Língua Portuguesa, Linguística e Literatura, Ano 05, n.11. 2009. Disponível em: https://ojs.ifsp.edu.br/magna/issue/view/148. Acesso em: 6 jun. 2026.

MOREIRA FILHO, J. L. Python para Linguística de *Corpus*: guia prático. 1. ed. São Paulo. Edição do Autor, 2021.

NOTAS & INFORMAÇÕES. O Brasil escolheu o atraso. O Estado de S. Paulo, São Paulo, 10 maio 2026. Disponível em: https://www.estadao.com.br/opiniao/o-brasil-escolheu-o-atraso/. Acesso em: 13 out. 2026.

PÊCHEUX, M. O Discurso: estrutura ou acontecimento. Campinas, SP: Pontes, 1990.

ISSN 2317-1316

RIBEIRO, Bruno. Atrito entre Kassab e Tarcísio esvazia apoio do PSD no interior paulista. Folha de S. Paulo, São Paulo, 18 maio 2026. Disponível em: https://www1.folha.uol.com.br/poder/2026/05/atrito-entre-kassab-e-tarcisio-esvazia-apoio-do-psd-no-interior-paulista.shtml. Acesso em: 13 out. 2026.

TRAQUINA, Nelson. Teorias do jornalismo: porque as notícias são como são. 2. ed. Florianópolis: Insular, 2005.