Além das Palavras: Detectando Técnicas Persuasivas em Notícias e Artigos de Opinião da Língua Portuguesa com BERT-Tiny

Adriel F. Trajano[1], **Erlon L. Avelino**[1], **Kamily A. de Oliveira**[1], **Yuri Almeida**[1]

[1]Centro de Informática – Universidade Federal da Paraíba (UFPB)
João Pessoa, PB, Brasil

{adriel.ferreira08, erlonlacerda1, kamilyfacul}@gmail.com, yuri@ci.ufpb.br

Abstract. *This paper presents an approach for the automatic identification of persuasion techniques in Portuguese. To this end, a dataset was automatically labeled using Large Language Models (LLMs) across eight categories, comprising seven persuasion techniques derived from SemEval23-T3 and one neutral class. Based on this dataset, the BERT-Tiny model was fine-tuned to establish a compact baseline for the classification task. Preliminary results achieved 86% accuracy under limited hardware conditions (NVIDIA T4 GPU), indicating a viable alternative for developing Portuguese-language resources for the analysis of persuasive strategies in disinformation.*

Resumo. *O presente trabalho propõe uma abordagem para a identificação automática de técnicas de persuasão em textos da língua portuguesa. Para isso, foi construída uma base de dados rotulada com Large Language Models (LLMs) em oito categorias, sendo sete técnicas de persuasão derivadas da SemEval23-T3 e uma classe neutra. A partir da base, foi realizado o fine-tuning do modelo BERT-Tiny, visando estabelecer uma baseline compacta para a tarefa de classificação. Os resultados preliminares alcançaram 86% de acurácia em ambiente de hardware limitado (GPU NVIDIA T4), indicando uma alternativa viável para o desenvolvimento de recursos de análise de estratégias persuasivas em desinformação para a língua portuguesa.*

1. Introdução

A disseminação da desinformação tornou-se um dos principais desafios contemporâneos, representando risco social, político e à saúde pública em escala global [World Economic Forum 2024]. No Brasil, esse fenômeno é especialmente relevante no campo político-eleitoral, impulsionado pelo uso intenso de redes sociais e pela polarização do debate público [Resende et al. 2019]. Esse contexto favorece a circulação de boatos e conteúdos enganosos, além de ampliar preocupações com a confiança nas notícias digitais, como aponta o *Digital News Report 2024* [Newman et al. 2024].

Grande parte das pesquisas sobre desinformação trata o problema como classificação binária, distinguindo conteúdos apenas como falsos ou verdadeiros [Oshikawa et al. 2020]. Essa abordagem, porém, é limitada: *fake news* abrange informação incorreta, desinformação e má-informação [Wardle and Derakhshan 2017], além de não indicar quais recursos discursivos favorecem a adesão do leitor. Por isso, é essencial analisar técnicas de persuasão para compreender *como* o conteúdo influencia o leitor

[Da San Martino et al. 2019, Piskorski et al. 2023], já que apelos emocionais, exageros e ataques pessoais podem tornar convincente uma narrativa distorcida ou manipulada.

Diante desse cenário, este trabalho propõe uma abordagem acessível para identificar automaticamente estratégias persuasivas em textos jornalísticos em língua portuguesa. Neste estudo, a noção de baixo custo refere-se a três dimensões: a redução do custo computacional, por meio do uso de um modelo compacto; a redução do custo de infraestrutura, ao realizar os experimentos em ambiente com GPU NVIDIA T4; e a redução do custo de anotação, ao empregar um *LLM* para rotular automaticamente as sentenças. Desenvolvida inicialmente como um projeto da disciplina de Processamento de Linguagem Natural (PLN), do Bacharelado em Ciência de Dados e Inteligência Artificial da UFPB, a pesquisa envolveu a construção de uma base de dados rotulada em sete categorias derivadas da Tarefa 3 da SemEval 2023 (SemEval23-T3) [Piskorski et al. 2023], além de uma categoria destinada a sentenças neutras. Em seguida, realizou-se o *fine-tuning* do modelo BERT-Tiny como *baseline* para a classificação dessas técnicas em português. Assim, as principais contribuições do estudo são a criação de um recurso em português e uma estratégia metodológica de baixo custo para anotação e classificação automática, visando apoiar novas pesquisas voltadas ao enfrentamento da desinformação.

2. Trabalhos Relacionados

A identificação automática de técnicas de persuasão tem sido explorada em PLN em espaços como o SemEval23-T3 [Piskorski et al. 2023] onde foram definidas 23 técnicas para detecção de persuasão em notícias multilíngues. Trabalhos anteriores também destacam que esse tipo de análise torna a detecção de propaganda mais interpretável [Da San Martino et al. 2019].

No Brasil, pesquisas como o Fake.Br Corpus investigam a classificação de notícias falsas em português [Monteiro et al. 2018], mas ainda se concentram exclusivamente na distinção entre falso e verdadeiro. O presente estudo avança nessa lacuna ao trazer uma abordagem focada em classificar as técnicas persuasivas de casos com viés politíco em português do Brasil.

3. Dados e Metodologia

Esta seção apresenta o percurso metodológico adotado no estudo, abrangendo a composição da base de dados a partir de diferentes fontes jornalísticas, os procedimentos de coleta e pré-processamento dos textos, a rotulação automática das sentenças e o treinamento do classificador BERT-Tiny.

3.1. Conjunto de dados

Para compor o conjunto foram reunidos quatro fontes, cobrindo notícias falsas, checagem de fatos, artigos de opinião e um recurso multilíngue usado como referência:

- **Conexão Política**: 1882 artigos de opinião sobre política e economia;
- **FakeTrueBr**: 1747 notícias em português rotuladas como falsas e com as suas respectivas notícias verdadeiras atreladas a elas;
- **G1 Fato ou Fake**: 3668 notícias coletadas via *web scraping* da seção de checagem de fatos do portal G1;
- **SemEval23-T3**: 529 textos em inglês rotulados com 23 técnicas de persuasão, utilizados exclusivamente como referência para a construção do *prompt* de rotulação automática.

3.2. Pré-processamento e rotulação

Coleta. A coleta dos textos foi conduzida conforme as características de cada fonte. No portal Conexão Política, foi realizado a raspagem das seções de análise e opinião, com identificação dos links das matérias, extração do conteúdo e recuperação de metadados. No Portal G1, a coleta partiu do *sitemap* contendo os links de prefixo *Fato ou Fake*, seguida da captura das páginas e da extração dos elementos textuais. Enquanto os conjuntos FakeTrueBr e SemEval23-T3 foram coletados a partir do [Monteiro et al. 2018], [Piskorski et al. 2023], respectivamente.

Pré-processamento. Após a coleta, as notícias foram segmentadas em sentenças, dado que as técnicas de persuasão tendem a aparecer mais especificamente em trechos. O corpus final contém 19.581 sentenças. Como os dados em português não possuíam rótulos de persuasão, foi realizada uma rotulação automática com *gpt-4.1-nano*, seguindo estudos recentes sobre *LLM*s como anotadores [Tan et al. 2024, Kasner et al. 2025]. O *prompt* incluiu instruções, descrição das categorias e exemplos *few-shot* do SemEval23-T3, usados apenas para orientar o rotulador.

Rotulação. Cada sentença foi associada a uma das oito categorias, além de receber uma justificativa gerada pelo *gpt-4.1-nano* para a atribuição do rótulo. Para manter compatibilidade, os rótulos foram armazenados em inglês e são descritos no texto em português: *Loaded_Language* como **Palavras fortes**, isto é, termos com alta carga emocional; *Name_Calling-Labeling* como **Imposição**, uso de rótulos ou julgamentos categóricos; *Doubt* como **Descredibilidade**, questionamento da credibilidade de pessoas, instituições ou fontes; *Repetition* como **Repetição**, recorrência de termos ou ideias; *Appeal_to_Fear-Prejudice* como **Apelo ao medo**, uso de ameaças ou riscos; *Flag_Waving* como **Apelo patriótico**, mobilização de identidade nacional ou coletiva; *Exaggeration-Minimisation* como **Exagero ou minimização**, amplificação ou redução da gravidade de fatos; e *No_Label* como **Neutro**, aplicado a sentenças que não apresentam nenhuma das técnicas persuasivas, incluindo, em geral, sentenças predominantemente informativas ou descritivas.

Sentença	**Rótulo**	**Justificativa**
"O presidente Jair Bolsonaro teria economizado 'um valor absurdamente maior' em viagem a Davos, em relação à ex-presidente Dilma Rousseff."	Exagero ou minimização	A expressão *valor absurdamente maior* sugere uma comparação exagerada para enfatizar a diferença de gastos.
"Se esses dois desgraçados que citei tivessem sido executados no primeiro crime, duas crianças estariam vivas em sua plenitude."	Palavras fortes	O termo *desgraçados* evoca forte reação emocional negativa, visando influenciar a percepção do leitor sobre as pessoas descritas.
"Não vamos cair na esparrela, na armadilha que o pessoal tá armando aí para fraudar as urnas."	Apelo ao medo	O texto insinua uma conspiração contra os eleitores, utilizando termos como *armadilha* e *fraudar as urnas* para gerar desconfiança e medo em relação ao processo eleitoral.

Tabela 1. Exemplos de sentenças com rótulos e justificativas

3.3. Treinamento do classificador

O treinamento foi conduzido por meio do *fine-tuning* do BERT-Tiny, uma variante compacta do modelo BERT [Devlin et al. 2019]. A escolha desse modelo se justifica pelo objetivo do trabalho de estabelecer uma *baseline* para a classificação de técnicas persuasivas, priorizando uma arquitetura mais leve e adequada a cenários com restrição de hardware.

A tarefa foi formulada como um problema de classificação multiclasse, no qual cada sentença tenha que ser associada a uma das oito categorias de persuasão definidas na etapa de rotulação. A base de dados foi dividida em 70% para treinamento, 15% para validação e 15% para teste, preservando a distribuição das classes em cada partição. O conjunto de treinamento foi utilizado para o ajuste dos parâmetros do modelo, o conjunto de validação para acompanhar o comportamento durante as épocas e o conjunto de teste para a avaliação de métricas.

Os experimentos foram executados no Google Colab com GPU NVIDIA T4. O modelo foi treinado por 20 épocas, com taxa de aprendizado de 5×10^{-5}, *warmup* de 0,1 e *weight decay* de 0,01. Esses hiperparâmetros foram adotados para manter o processo de treinamento estável e compatível com a proposta computacionalmente acessível.

4. Resultados

Os resultados preliminares foram avaliados por meio das curvas de perda, matriz de confusão, relatório de classificação e exemplos qualitativos. No conjunto de teste, o modelo alcançou 86% de acurácia indicando desempenho consistente entre as diferentes categorias. A Figura 1 mostra que o modelo reduziu a perda nas épocas iniciais e estabilizou ao final do treinamento, especialmente entre as épocas 19 e 20. Embora exista diferença entre treino e validação, o comportamento geral indica convergência sem instabilidade nas últimas épocas.

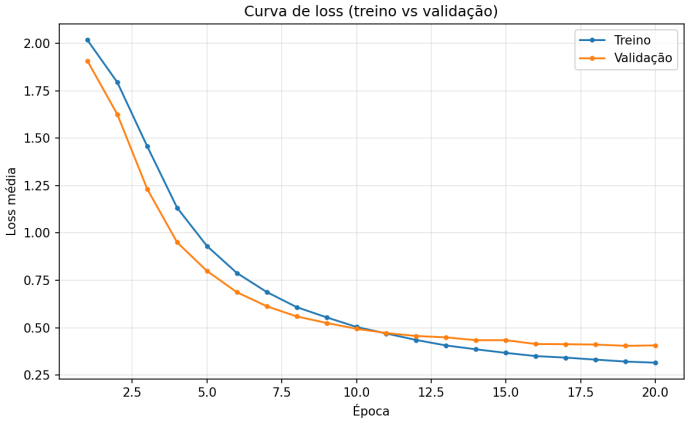

Figura 1. Curvas de perda nos conjuntos de treinamento e validação.

Pela matriz de confusão (Figura 2) podemos perceber que classes semanticamente próximas se confundem com frequência, como palavras fortes, apelo patriótico e exagero. Enquanto a tabela de resultados de classificação (Tabela 2) sugere que a classe neutro obteve dificuldade em separar sentenças sem persuasão explícita de trechos com marcas discursivas sutis, diminuindo sua precisão.

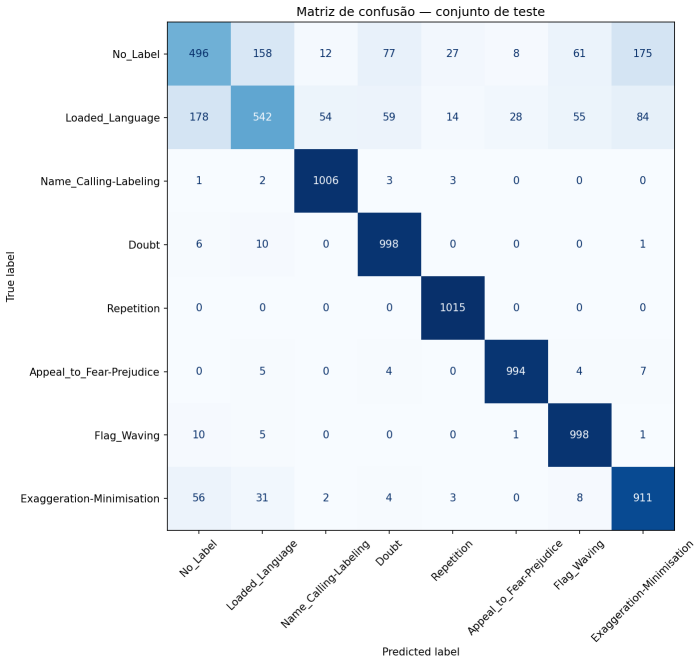

Figura 2. Matriz de confusão no conjunto de teste.

Class	Precision	Recall	F1-score	Support
Neutro	0.65	0.51	0.57	1015
Palavras fortes	0.70	0.53	0.60	1015
Imposição	0.95	1.00	0.97	1015
Descredibilidade	0.90	0.98	0.94	1014
Repetição	0.96	1.00	0.98	1014
Apelo ao medo	0.97	0.98	0.97	1015
Apelo patriótico	0.89	0.98	0.93	1015
Exagero ou minimização	0.78	0.90	0.84	1014

Tabela 2. Métricas por classe no conjunto de teste.

A análise qualitativa reforça essas limitações. No exemplo 1, a sentença foi classificada como apelo patriótico, sugerindo que o modelo pode associar referências a eleições, poder político e regimes autoritários a categorias identitárias, mesmo quando o rótulo esperado está ligado à ”Palavras Fortes”.

> **Exemplo 1:** “Mais sério que o desvio foi o propósito do mesmo: comprar apoios, bancar eleições e se perpetuar no poder, além de ajudar a ditaduras amigas.”
> **Rótulo esperado:** Palavras fortes $\longrightarrow$ **Predição:** Apelo patriótico

No exemplo 2, uma sentença originalmente neutra foi classificada como exagero ou minimização. Esse caso indica que expressões relacionadas à duração de um governo ou à continuidade política podem ser interpretadas pelo modelo como intensificação discursiva, mesmo quando aparecem em contexto informativo.

Exemplo 2: "Isso não significa que as negociações estejam concluídas e que Israel em breve obterá um novo governo que encerrará o reinado de 10 anos de Netanyahu."
Rótulo esperado: Neutro $\longrightarrow$ **Predição:** Exagero ou minimização

Também houve casos em que o modelo identificou corretamente a técnica persuasiva. No exemplo 3, a menção ao período eleitoral e ao contexto de mobilização política foi classificada como *Flag Waving*, categoria associada ao apelo patriótico.

Exemplo 3: "O período de segundo turno das eleições gerais..."
Rótulo esperado: Apelo patriótico $\longrightarrow$ **Predição:** Apelo patriótico

No exemplo 4, outro acerto ocorreu em uma sentença com rotulação depreciativa explícita. O uso de termos como "bandido é lixo" caracteriza *Name Calling-Labeling*, pois atribui um rótulo negativo para produzir julgamento moral e reação emocional.

Exemplo 4: "Será que é preciso que você passe por isso pra entender que bandido é lixo e deve ser extirpado da face da terra?"
Rótulo esperado: Imposição $\longrightarrow$ **Predição:** Imposição

Em síntese, modelos compactos conseguem aprender padrões associados às técnicas de persuasão mesmo com rotulação automática, mas categorias próximas e sentenças neutras continuam sendo os principais desafios. Esses exemplos também indicam que a tarefa exige mais do que reconhecer palavras isoladas: é necessário considerar o contexto discursivo e a função retórica de cada trecho.

5. Conclusão

Este trabalho apresenta uma abordagem de baixo custo para a identificação automática de técnicas de persuasão em português. Para isso, foi construída uma base de dados com 19.581 sentenças rotuladas em oito categorias, incluindo sete técnicas persuasivas derivadas da SemEval23-T3 e uma classe neutra. Além disso, foi documentado um procedimento de anotação automática com justificativas geradas pelo *gpt-4.1-nano* e realizado o *fine-tuning* do BERT-Tiny como *baseline* compacta para a tarefa.

Ainda assim, os resultados desse estudo devem ser interpretados como exploratórios, pois a dependência de rótulos gerados por *LLM*s pode introduzir inconsistências, especialmente em categorias semanticamente próximas e em sentenças neutras.

Para trabalhos futuros, sugere-se realizar a validação humana das anotações, comparar diferentes estratégias de rotulação e modelos de linguagem, além de investigar abordagens *multilabel* para identificação de múltiplas técnicas persuasivas nas sentenças.

6. Declaração de uso de ferramentas de Inteligência Artificial

Durante a preparação deste trabalho, foi utilizado o ChatGPT com a finalidade de auxiliar na revisão gramatical do texto, na organização das ideias e no aprimoramento da clareza da redação. Após a utilização, o conteúdo foi revisado e editado pelos próprios autores, que assumem total responsabilidade por todas as informações apresentadas no artigo publicado.

Referências

Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., and Nakov, P. (2019). Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5636–5646.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kasner, Z., Zouhar, V., Schmidtová, P., Kartáč, I., Onderková, K., Plátek, O., Gkatzia, D., Mahamood, S., Dušek, O., and Balloccu, S. (2025). LLMs as span annotators: A comparative study of LLMs and humans. *arXiv preprint arXiv:2504.08697*.

Monteiro, R. A., Santos, R. L. S., Pardo, T. A. S., Almeida, T. A., Ruiz, E. E. S., and Vale, O. A. (2018). Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *Proceedings of the 13th International Conference on Computational Processing of the Portuguese Language*, pages 324–334.

Newman, N., Fletcher, R., Robertson, C. T., Ross Arguedas, A., and Nielsen, R. K. (2024). Reuters institute digital news report 2024. Technical report, Reuters Institute for the Study of Journalism, University of Oxford.

Oshikawa, R., Qian, J., and Wang, W. Y. (2020). A survey on natural language processing for fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France. European Language Resources Association.

Piskorski, J., Stefanovitch, N., Da San Martino, G., and Nakov, P. (2023). SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Resende, G., Melo, P., Sousa, H., Messias, J., Vasconcelos, M., Almeida, J. M., and Benevenuto, F. (2019). (mis)information dissemination in whatsapp: Gathering, analyzing and countermeasures. In *Proceedings of The Web Conference 2019*, pages 818–828.

Tan, Z., Beigi, A., Wang, S., Guo, R., Bhattacharjee, A., Jiang, B., Karami, M., Li, J., Cheng, L., and Liu, H. (2024). Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.

Wardle, C. and Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. Technical report, Council of Europe.

World Economic Forum (2024). The global risks report 2024. Technical report, World Economic Forum.