Heurísticas vs. LLMs na triagem de *tweets*: avaliando pistas linguísticas para tarefas de PLN em português

Laura Pessine Teixeira[1], **Helena de Medeiros Caseli**[1]

[1]Departamento de Computação – Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – 13565-905 – São Carlos – SP – Brasil

laurapessine@estudante.ufscar.br, helenacaseli@ufscar.br

Abstract. *This paper presents an intelligent triage pipeline to pre-process informal texts for Natural Language Processing (NLP) tasks. The pipeline scores text utility based on linguistic clues from the rhetorical dimension of argumentation (clarity, credibility, emotional appeal, and organization). An intrinsic evaluation compared the proposed pipeline with a Large Language Model (LLM) against a human-annotated gold standard. Results show the pipeline excels in objective features, while subjective elements require deep contextual models. These results indicate that a hybrid architecture may be the most effective strategy for textual data curation.*

Resumo. *Este artigo apresenta um pipeline de triagem inteligente para pré-processamento de textos informais em tarefas de Processamento de Linguagem Natural (PLN). O pipeline pontua a utilidade textual usando pistas linguísticas fundamentadas na dimensão retórica da argumentação (clareza, credibilidade, apelo emocional e organização). Uma avaliação intrínseca comparou o pipeline proposto com um Large Language Model (LLM) utilizando um padrão ouro humano. Os resultados mostram que o pipeline obteve maior precisão em características objetivas, enquanto elementos subjetivos exigem modelos contextuais. Esses resultados indicam que uma arquitetura híbrida pode ser uma estratégia mais eficaz para a curadoria de dados textuais.*

1. Introdução

As mídias sociais, particularmente o Twitter/X, consolidaram-se como fontes de dados relevantes para a compreensão de dinâmicas sociopolíticas. O volume contínuo de informações nessa plataforma estabelece um ecossistema textual de processamento complexo, onde a seleção de conteúdos adequados para tarefas de Processamento de Linguagem Natural (PLN) representa um desafio técnico significativo. Embora a área de PLN forneça a base metodológica para a análise desses textos [Caseli et al. 2024], a eficácia das aplicações depende da qualidade dos dados de entrada. No domínio político brasileiro – contexto no qual este trabalho se insere –, as publicações caracterizam-se por discurso polarizado e informalidade, exigindo curadoria prévia para atenuar o ruído e assegurar a precisão de tarefas como sumarização e análise de sentimentos.

Este artigo propõe um *pipeline* de triagem inteligente para pré-processamento que pontua pistas linguísticas retóricas para ranquear a utilidade de cada *tweet*, ou seja, sua adequação a tarefas específicas de PLN. Por exemplo, *tweets* ricos em autoridades e termos técnicos favorecem a sumarização, enquanto o uso de *emojis* e caixa alta beneficia a

análise de sentimentos. Essa curadoria baseada em utilidade pode melhorar a qualidade do *dataset*, gerando modelos mais precisos e promovendo explicabilidade [Bowman 2024]. A metodologia é validada contrastando o *pipeline* baseado em heurísticas com um *Large Language Model* (LLM) frente a um *gold standard* humano.

2. Fundamentação teórica

A avaliação automática da qualidade da argumentação quantifica a eficácia de um discurso em linguagem natural. A base teórica deste trabalho apoia-se na taxonomia de [Wachsmuth et al. 2017], focando na dimensão retórica, que demonstra maior viabilidade computacional em textos curtos por abranger características de superfície extraíveis de forma sistemática por ferramentas de PLN.

A adaptação dessa taxonomia para a política brasileira [Silva 2023] definiu pistas linguísticas observáveis para os aspectos retóricos de [Wachsmuth et al. 2017]. Este trabalho implementa a extração dessas pistas no seguinte mapeamento: (i) **Credibilidade**: presença de dados quantitativos ou temporais, termos técnicos e autoridades; (ii) **Apelo emocional**: *emojis*, caixa alta, *hashtags*, pronomes em 1ª pessoa e repetições de pontuação (!/?); e (iii) **Clareza**: penalizada por desvios ortográficos e complexidade sintática.

A Inteligência Artificial Centrada em Dados (*Data-Centric AI*) enfatiza que o desempenho em PLN depende da melhoria contínua dos *datasets* [Chen 2022], onde a filtragem estratégica reduz custos e mitiga vieses [Albalak et al. 2024]. É essencial, contudo, distinguir o "ruído prejudicial" do "ruído útil", que carrega intencionalidade semântica (como pontuações de sentimento) [Al Sharou et al. 2021]. Ao avaliar critérios isolados para mensurar a relevância global [Pinto et al. 2017], o *pipeline* proposto gera uma pontuação objetiva, consolidando a curadoria como etapa indispensável para adequar os textos às tarefas subsequentes de IA.

3. Metodologia

3.1. *Corpus* de estudo

Utilizou-se uma amostra de 100 tweets do *Interfaces Twitter Elections Dataset* (ITED-Br), composto por cerca de 282 milhões de *tweets* coletados nas eleições presidenciais brasileiras de 2022 [Iasulaitis et al. 2025]. A escolha justifica-se pela presença de disputas narrativas que geram linguagem polarizada, dificultando o processamento automático por modelos de PLN treinados em textos formais.

3.2. *Pipeline* de filtragem inteligente

Com base na fundamentação teórica [Wachsmuth et al. 2017, Silva 2023], desenvolveu-se um *pipeline* implementado em Python que inicializa os *scores* retóricos em zero e os atualiza iterativamente. Para os aspectos de credibilidade e apelo emocional, soma-se 1 ponto a cada ocorrência isolada de suas respectivas pistas (Tabela 1). A clareza, por sua vez, é calculada por meio de acréscimos e penalidades, como detalhado a seguir.

Na prática, o *pipeline* emprega a biblioteca *spaCy*[1] (modelo `pt_core_news_lg`) para a marcação morfossintática e o Reconhecimento de Entidades Nomeadas (NER),

[1] `https://spacy.io`

Tabela 1. Relação das pistas linguísticas com os *scores* dos aspectos.

Pistas linguísticas	Credibilidade	Apelo emocional	Clareza
Autoridade/Dado, *Hashtag*, Termo técnico	+1	—	—
Emojis, Caixa alta, Repetição de pontuação	—	+1	—
Primeira pessoa/Relato pessoal	+1	+1	—
Erros de língua portuguesa	—	—	-1
Métricas textuais (NILC-Metrix)	—	—	+1 a -2

em que as entidades válidas extraídas computam diretamente a pista de autoridade/dado. Para a identificação de termos técnicos, as palavras do *tweet* são filtradas e contrastadas com o léxico geral PortiLexicon-UD[2] [Lopes et al. 2022] e as palavras ortograficamente corretas que estão ausentes nesse mapeamento de uso cotidiano são classificadas como vocabulário especializado.

O cálculo da clareza engloba processamento sintático e semântico adicional. Os erros de língua portuguesa são detectados via *pyspellchecker*[3], e cada desvio subtrai 1 ponto da pontuação. Em paralelo, a API do NILC-Metrix[4] [Leal et al. 2024] extrai 11 métricas sintáticas e semânticas do texto (incluindo índices de complexidade e legibilidade). Tais métricas são submetidas a limiares empíricos no algoritmo que penalizam o texto (subtraindo até 2 pontos em casos de alta prolixidade ou baixa coesão) ou o bonificam (+1 ponto).

4. Resultados e discussão

A avaliação intrínseca comparou o *pipeline* proposto com predições de um LLM (GPT). O *prompt* forneceu definições e exemplos das pistas, exigindo a extração dos trechos exatos em CSV. Como *gold standard*, dois avaliadores anotaram a saída ideal para uma amostra aleatória de 100 *tweets* (balanceados entre os candidatos), extraídos do *corpus* original sem filtros prévios por usuário ou período. A avaliação consistiu em verificar qual abordagem (heurística ou generativa) mais se aproximou dessa referência humana.

Para mitigar a subjetividade de pistas como autoridade e termo técnico, 20 instâncias foram analisadas em intersecção. As divergências geraram regras estritas: excluíram-se siglas e inícios de sentenças da caixa alta, classificaram-se neologismos e gírias como erros ortográficos e validaram-se como autoridades apenas figuras públicas, siglas partidárias e coletivos no sujeito. As 80 instâncias restantes foram anotadas independentemente sob essas métricas.[5] Os resultados (Tabela 2) evidenciam vantagens distintas conforme a natureza da pista.

Nas pistas objetivas, o *pipeline* igualou ou superou o LLM. Destaca-se a eficácia na identificação de caixa alta, onde o modelo generativo falhou ao ignorar maiúsculas acentuadas. O *pipeline* também mostrou maior aderência metodológica nas repetições de sinais de pontuação, limitando-se à interrogação e exclamação, enquanto o LLM in-

[2]https://github.com/LuceleneL/PortiLexicon-UD

[3]https://github.com/barrust/pyspellchecker

[4]https://github.com/nilc-nlp/nilcmetrix

[5]Devido ao foco na resolução de divergências por consenso para a construção do *gold standard*, não houve cálculo formal de métricas de concordância interanotadores.

Tabela 2. Comparativo de performance entre pipeline e LLM por pista linguística.

Pistas linguísticas	Categoria	Método de maior performance
Emojis, *Hashtags*, Pronomes em 1ª pessoa	Objetiva	Empate
Caixa alta, Repetição de pontuação	Objetiva	***Pipeline***
Autoridades, Termos técnicos	Subjetiva	Empate
Erros de língua portuguesa	Subjetiva	***Pipeline***

cluiu erroneamente sinais sem apelo emocional intenso, como reticências. Além disso, o *pipeline* identificou um volume superior de abreviações informais e erros ortográficos.

As pistas de caráter subjetivo expuseram limitações em ambas as abordagens. Na identificação de autoridades, observou-se um empate com perfis complementares: o LLM foi mais eficaz em entidades institucionais formais (e.g., "TSE" e "forças armadas"), enquanto o *pipeline* destacou-se em referências informais e geográficas com uso político (e.g., "América Latina" e "nine"). A detecção de termos técnicos foi a mais complexa: o LLM demonstrou permissividade ao classificar palavras comuns como especializadas, ao passo que o *pipeline* gerou falsos positivos com gírias políticas.

5. Conclusão e trabalhos futuros

O *pipeline* de triagem inteligente apresentado demonstrou ser uma etapa de pré-processamento eficaz para mensurar objetivamente a utilidade de *tweets* para PLN, fundamentado em pistas linguísticas retóricas. A análise revelou que o emprego de uma arquitetura híbrida (unindo regras explícitas a interpretações contextuais) pode ser a abordagem mais promissora para a curadoria de textos informais.

Como trabalhos futuros, propõe-se a integração de recursos complexos, como vetores de palavras (*word embeddings*), especificamente o modelo Wang2Vec [Hartmann et al. 2017], para detecção semântica de fatos históricos, e o uso do *Toxic Language Dataset for Brazilian Portuguese* (ToLD-Br) [Leite et al. 2020] para treinar classificadores que aprimorem a medição do apelo emocional negativo, ampliando a cobertura do sistema. O código-fonte do pipeline encontra-se disponível publicamente em: `https://github.com/LALIC-UFSCar/ic-triagem-tweets-politica`.

Agradecimentos

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) via Programa Institucional de Bolsas de Iniciação Científica (PIBIC). Esta pesquisa integra o projeto "Análise de grandes volumes de dados políticos e redes complexas: mineração, modelagens e aplicações em Ciência Política Computacional" financiado pela FAPESP (#22/03090-0) e faz parte do Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR). Agradecemos também à Pró-Reitoria de Graduação (ProGrad) da UFSCar pelo auxílio financeiro concedido para a participação no evento. Em conformidade com as diretrizes do evento, os autores declaram o uso do modelo generativo de IA Gemini (Google) para auxiliar na revisão, formatação em LaTeX e aprimoramento da fluidez textual deste artigo. A responsabilidade integral pelo conteúdo, análises e conclusões permanece dos autores.

Referências

Al Sharou, K., Li, Z., and Specia, L. (2021). Towards a better understanding of noise in natural language processing. In Mitkov, R. and Angelova, G., editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 53–62, Held Online. INCOMA Ltd.

Albalak, A., Elazar, Y., Xie, S. M., Longpre, S., Lambert, N., Wang, X., Muennighoff, N., Hou, B., Pan, L., Jeong, H., Raffel, C., Chang, S., Hashimoto, T., and Wang, W. Y. (2024). A survey on data selection for language models.

Bowman, S. R. (2024). Eight Things to Know about Large Language Models. *Critical AI*, 2(2).

Caseli, H. d. M., Nunes, M. d. G. V., and Pagano, A. (2024). O que é PLN? In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, chapter 1. BPLN, 3 edition.

Chen, H. (2022). *Data Quality Evaluation and Improvement for Machine Learning*. PhD thesis, University of North Texas, USA.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, Uberlândia, Brazil.

Iasulaitis, S., Valejo, A. D. B., Greco, B. C., Perillo, V. G., Messias, G. H., Vicari, I., and Interfaces—Center for Sociopolitical Studies of Algorithms and Artificial Intelligence (2025). The Interfaces Twitter Elections Dataset: Construction process and characteristics of big social data during the 2022 presidential elections in Brazil. *PLOS ONE*, 20(2):e0316626.

Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., and Aluísio, S. M. (2024). NILC-Metrix: Assessing the Complexity of Written and Spoken Language in Brazilian Portuguese. *Language Resources and Evaluation*, 58(1):73–110.

Leite, J. A., Silva, D. F., Bontcheva, K., and Scarton, C. (2020). Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, China.

Lopes, L., Duran, M. S., Fernandes, P., and Pardo, T. A. S. (2022). PortiLexicon-UD: A Portuguese Lexical Resource According to Universal Dependencies Model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6635–6643, Marseille, France. European Language Resources Association.

Pinto, A., Gonçalo Oliveira, H., Figueira, Á., and Alves, A. O. (2017). Predicting the relevance of social media posts based on linguistic features and journalistic criteria. *New Generation Computing*, 35(4):451–472.

Silva, C. F. d. (2023). *Abordagem computacional para avaliação automática da qualidade da argumentação na dimensão retórica de tweets no domínio da política brasileira*. PhD thesis, Universidade Federal de São Carlos, São Carlos.

Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T. A., Hirst, G., and Stein, B. (2017). Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.