Tweets8JanSPI: Uma Base de Dados para Treinamento de Modelos de Mineração de Argumentos em Redes Sociais

Livia Alabarse dos Santos[1], Pedro Henrique Araujo Farias[1], Guilherme de Abreu Schulz[1], Renata Araujo[1]

[1]Faculdade de Computação e Informática (FCI)
Universidade Presbiteriana Mackenzie – São Paulo, SP – Brasil

{liviaalabarse.santos,pedrohenriquearaujo.farias,guilherme.schulz}
@mackenzista.com.br, renata.araujo@mackenzie.br

Abstract. *This work presents the creation of an annotated dataset of tweets related to the invasion of the headquarters of the Three Powers of the Republic in Brasília in 2023, focusing on the task of argument mining. The study describes the manual annotation process focused on three categories: irony, positioning, and sentiment, based on an annotation manual developed specifically for the context of this project. The dataset with 2,935 labeled tweets was subsequently subjected to an analysis of agreement between annotators, making it possible to identify the high degree of difficulty of activities related to natural language, essential for argument mining and understanding discussions on social networks.*

Resumo. *Este trabalho apresenta a criação de uma base de dados anotada com tweets relacionados à invasão à sede dos Três Poderes da República em Brasília em 2023, com foco na tarefa de mineração de argumentos. O estudo descreve o processo de anotação manual voltado para três categorias: ironia, posicionamento e sentimento, fundamentado em um manual de anotação desenvolvido especificamente para o contexto deste projeto. O dataset com 2935 tweets rotulados foi posteriormente submetido a uma análise de concordância entre anotadores, tornando possível identificar o alto grau de dificuldade de atividades relacionadas à linguagem natural, essenciais para a compreensão de discussões e a mineração de argumentos em redes sociais.*

1. Introdução

A evolução das redes sociais e a expansão das interações online têm tornado cada vez mais relevante o estudo de como os usuários discutem, argumentam e compartilham suas opiniões nessas plataformas. No contexto da Ciência de Dados e do Processamento de Linguagem Natural, a Mineração de Argumentos (MA) surge como uma abordagem promissora para identificar, extrair e analisar a estrutura argumentativa presente nesses textos [Stede e Scheineder 2022]. A aplicação da MA em conteúdos de redes sociais enfrenta desafios como a qualidade dos textos (carregados de erros ortográficos), a informalidade (abreviações, gírias e etc.) e poucas bases de dados anotadas, especialmente em português brasileiro [Bosc, Cabrio e Villata 2016][Silva *et al*. 2024].

O projeto HEIWA busca desenvolver uma plataforma para análise de discussões em redes sociais a partir da MA, com foco no cenário brasileiro [Santos *et al*. 2025]. A plataforma HEIWA está organizada em quatro principais eixos: curadoria, extração e limpeza, estruturação de argumentos e visualização de dados. O presente trabalho está inserido especificamente no eixo de curadoria do projeto, responsável pelo preparo de corpora em português brasileiro para a tarefa de MA.

Um componente relevante da MA é o tratamento da subjetividade dos textos [Stede e Schneider 2022]. No caso de conteúdos das redes sociais, essa subjetividade é ainda maior, dada a informalidade das falas e o pouco espaço de expressividade das postagens. Dentro da estratégia de MA do projeto HEIWA, entende-se que a detecção, em cada postagem, de elementos como sentimento, posição e ironia é fundamental para a detecção de estruturas argumentativas [Santos *et al*. 2025].

O objetivo deste trabalho é descrever o processo de construção e anotação de uma base com uma amostra de *tweets* em português brasileiro, extraídos da base Tweet_Eleições_2022 [Silva *et al.* 2024], relativos à invasão à sede dos Três Poderes em Brasília, em 8 de janeiro de 2023. O processo de organização e de anotação da base teve o intuito de prepará-la para o treinamento de modelos de aprendizado de máquina para a MA em conteúdos de redes sociais, considerando ironia, sentimento e posição.

2. Trabalhos Relacionados

Entre os trabalhos relacionados à anotação de dados para o treinamento de modelos, há predominância do emprego de *Large Language Models* (LLMs) para a classificação dos textos. Embora o uso de modelos para rotulamento dos dados confira escalabilidade ao processo, esse método não captura a subjetividade e as nuances sociais e culturais inerentes às discussões em redes sociais como a anotação manual, considerada adequada para a consolidação de uma base padrão ouro [Pustejovsky e Stubbs 2012].

Santos e Berton (2023) separaram um conjunto randomizado de 20% de uma amostra extraída do então Twitter a respeito das eleições presidenciais de 2022 no Brasil e classificaram manualmente os *tweets* de acordo com o sentimento expresso. Em seguida, utilizaram as anotações iniciais para treinar um modelo de classificação de sentimentos, ampliando o número de textos anotados. A anotação da base Tweet_Eleições_2022 realizada por Santos *et al.* (2025) apresentou uma abordagem híbrida, dividida em dois momentos: iniciando com anotação manual em uma oficina oferecida a alunos de graduação da área de computação e, em seguida, anotando cerca de 1000 *tweets* com o apoio de LLMs. Lima Filho *et al.* (2024) realizaram anotação totalmente automática, valendo-se de LLMs para analisar cerca de 2780 publicações em redes sociais, em inglês e português brasileiro, e classificá-las de acordo com a presença de sinais de depressão, resultando na base *DepressSet*.

Este trabalho se diferencia pela criação de uma base de 2935 *tweets* anotados de modo totalmente manual em três dimensões: análise de sentimentos, detecção de ironia e identificação de posicionamento, cujo processo de anotação foi apoiado por um manual elaborado pelos anotadores, de modo a apresentar definições claras e exemplos específicos do contexto dos dados a serem anotados.

3. Coleta dos dados e seleção da amostra

A base de dados Tweet_Eleições_2022 [Silva *et al.* 2024] foi criada a partir do consumo da API da plataforma Twitter durante o período das eleições presidenciais de 2022 no Brasil, com o intuito de capturar *tweets* relacionados às eleições e aos temas políticos relevantes. Para a composição da base de dados anotada neste trabalho, foi selecionado um recorte de cerca de 10 mil *tweets* do Tweet_Eleições_2022 para três períodos de

tempo distintos relacionados à invasão à sede dos Três Poderes, em 8 de janeiro: início (manhã), meio (tarde) e fim (noite) do evento, totalizando um conjunto inicial de 30 mil *tweets*. O recorte foi então randomizado para impedir que a ordem cronológica influenciasse o processo de anotação, de modo que fosse possível capturar uma maior variação discursiva, uma vez que a dinâmica das interações acerca da invasão tende a se modificar conforme o desenrolar dos acontecimentos [Bertanha e Araujo 2024].

Após a amostragem, realizou-se um processo de filtragem qualitativa, que considerou o tamanho dos *tweets*, dado que textos muito curtos tendem a oferecer pouco contexto e podem dificultar a interpretação de sentimento, posicionamento ou ironia. Por esse motivo, foram excluídos *tweets* com menos de 100 caracteres, garantindo que apenas mensagens com conteúdo suficiente permanecessem na base. Além disso, *tweets* altamente similares foram removidos para evitar duplicidades.

O conjunto resultante foi convertido para o formato JSON, compatível com a ferramenta de anotação selecionada, preservando a estrutura textual e as informações necessárias para posterior rotulagem.

4. Processo de anotação

De modo a orientar os anotadores durante o processo de classificação dos textos, foi desenvolvido um manual de anotação, o qual estabelece uma abordagem sistemática ao processo de anotação e, principalmente, garante a concordância entre os anotadores. O manual descreve a base de dados a ser anotada e contextualiza o anotador quanto ao seu objetivo, justificando sua existência como um guia prático para a mitigação de problemas comuns ao processo de anotação, muitas vezes causados pela subjetividade inerente à linguagem, desafio comum principalmente nas atividades de detecção de ironia e análise de sentimento [Hovy e Lavid 2010][Pustejovsky e Stubbs 2012].

O manual de anotação apresenta a definição de cada uma das três dimensões a serem anotadas e lista uma série de exemplos de *tweets* corretamente anotados, os quais devem ser tomados como referência. Além disso, exibe problemas que os anotadores podem identificar durante a classificação de *tweets*, propondo que essas situações sejam minuciosamente analisadas por cada um dos participantes do processo. A anotação teve como objetivo a construção de um conjunto de dados destinado ao treinamento de um modelo de MA a partir do rotulamento de textos extraídos de *tweets* com base em três das tarefas consideradas essenciais para a compreensão dos discursos em redes sociais: análise de sentimentos, identificação de posicionamento e detecção de ironia [Santos *et al.* 2025]. Para esse fim, foram estabelecidos rótulos específicos para cada dimensão analisada: na dimensão de ironia, “sim” (contém ironia) e “não” (não contém ironia); na dimensão de sentimento, “positivo”, “negativo” e “neutro”; e, na dimensão de posicionamento, “a favor”, “contra” e “neutro”.

Como ferramenta de anotação, foi selecionado o **Doccano**, uma ferramenta de código aberto para anotação colaborativa de dados [Nakayama *et al*. 2018]. O *software* permite a importação de bases de dados para anotação, bem como configuração de rótulos e validação de anotações. Após a preparação, filtragem da amostra e configuração do ambiente de anotação, deu-se início à etapa de anotação, adotando abordagem manual em detrimento da utilização de LLMs de modo a garantir a alta

qualidade da anotação [Fuchs *et al.* 2026], estabelecendo uma base padrão ouro validada por humanos [Pustejovsky e Stubbs 2012], a qual posteriormente poderá ser utilizada no treinamento de modelos.

Considerando o grande volume de dados na amostra (30 mil *tweets*), e levando em conta o processo demorado de anotação manual, optou-se por anotar manualmente uma amostra reduzida de até 3 mil *tweets*. Essa decisão teve como base a viabilidade operacional e profundidade analítica, uma vez que o processo de anotação manual demanda tempo, atenção e consistência.

A anotação foi conduzida pelos autores seguindo o manual padronizado de anotação desenvolvido previamente, com uma média de 100 *tweets* anotados diariamente ao longo de um período de dois meses. Cada *tweet* foi lido integralmente e classificado conforme o contexto discursivo. Nos casos de dúvida, os anotadores utilizaram um grupo de discussão, no qual eram debatidos trechos ambíguos e de difícil interpretação. Essas revisões coletivas permitiram ajustar os critérios e reduzir divergências, fortalecendo a coerência interna das anotações.

Durante o processo, foram observadas dificuldades na identificação de ironia, especialmente em textos curtos ou com uso de sarcasmo implícito, o que exigiu discussões adicionais e refinamento dos critérios. Após o término da anotação, os dados foram exportados como um arquivo JSON para cada anotador, contendo os rótulos correspondentes a cada categoria.

4.1. Análise da anotação

Ao todo, foram manualmente anotados 2935 *tweets* em cada uma das três categorias. Realizou-se uma análise de concordância entre anotadores a fim de identificar o grau de alinhamento nos rotulamentos de cada categoria. As comparações foram conduzidas de forma par-a-par para cada uma das três dimensões, de modo a identificar padrões de divergência entre os anotadores, bem como categorias com maior ou menor consenso e as classificações predominantes de cada dimensão (Figura 1).

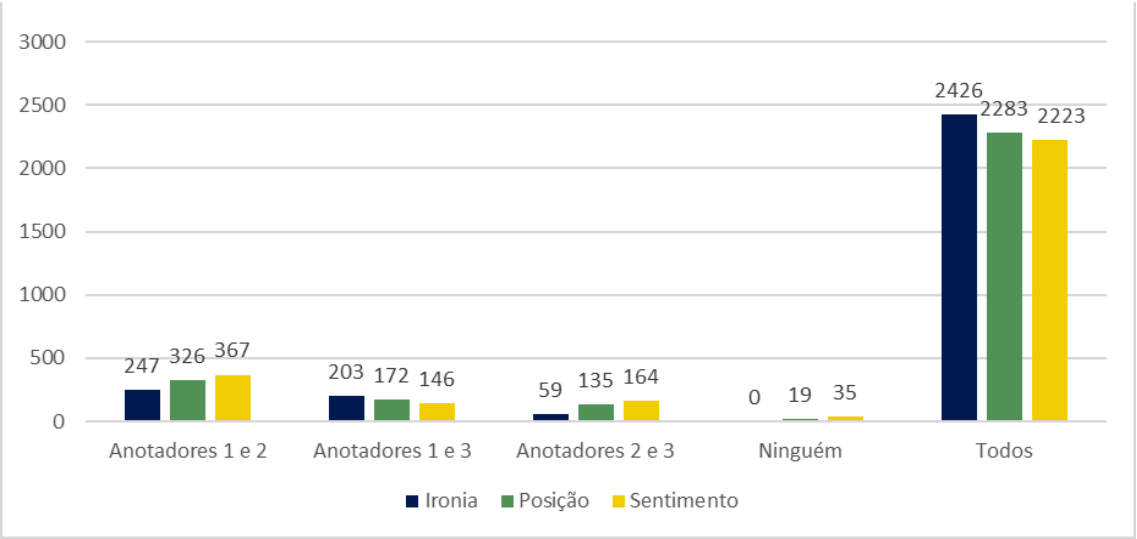

Figura 1. Concordância entre anotadores

No gráfico da Figura 1, o primeiro bloco indica a quantidade de *tweets* em que somente os anotadores 1 e 2 concordam entre si. O mesmo vale para o segundo e terceiro bloco, em que somente os anotadores 1 e 3 e os anotadores 2 e 3 concordam entre si, respectivamente. O penúltimo bloco indica em quantos *tweets* nenhum dos anotadores concorda entre si. Por fim, o último bloco denota a quantidade de *tweets* em que todos os anotadores concordam entre si. Nota-se que os anotadores apresentaram boa concordância, onde em um total de 2935, o número de *tweets* em que todos os três anotadores concordaram entre si manteve-se acima de 2200.

Quando os três anotadores discordam entre si, como é possível observar nas tarefas de análise de sentimento e detecção de posicionamento (denotados no gráfico por "Ninguém"), torna-se necessário o rotulamento dos textos pendentes por um quarto anotador. Após anotação pelo quarto anotador, obteve-se o resultado final das anotações:

Tabela 1. Resultado final das anotações

	Ironia		**Sentimento**			**Posicionamento**		
Rótulo	Não	Sim	Negativo	Neutro	Positivo	Contra	Neutro	Favor
Qtd.	2711	224	2261	430	244	1942	505	488

Observa-se que o resultado foi uma base desbalanceada, com alto número de *tweets* de sentimento negativo e posicionamento contra a invasão da sede dos Três Poderes. Ainda que tenha sido adotada uma estratégia para capturar maior diversidade de discursos, o desbalanceamento era esperado em razão da natureza do acontecimento estudado. Quanto às métricas de concordância entre anotadores, foram calculados o Kappa de Fleiss e o Alpha de Krippendorff para cada uma das dimensões anotadas:

Tabela 2. Métricas de concordância entre anotadores

	Métrica de concordância	
Atributo	**Kappa de Fleiss**	**Alpha de Krippendorff**
Sentimento	0,59	0,63
Posicionamento	0,71	0,64
Ironia	0,38	0,38

Obteve-se concordância moderada para sentimento e concordância substancial para posicionamento de acordo com os Kappas de Fleiss. Segundo os valores do Alpha de Krippendorff, tanto sentimento quanto posicionamento apresentam confiabilidade de concordância baixa. O desbalanceamento das categorias anotadas pode impactar os coeficientes de concordância, uma vez que as métricas são sensíveis à distribuição desigual das classes [Artstein e Poesio 2008][Krippendorff 2019][Warrens 2010]. Os baixos índices ressaltam a complexidade dos discursos políticos e a dificuldade da tarefa de anotação de textos relacionados a um evento extremamente sensível e atípico.

5. *Dataset* anotado

Os metadados da base de dados anotada estão descritos na Tabela 3:

Tabela 3. Metadados da base de dados Tweets8JanSPI

Campo	Tipo	Descrição
conversation_id	*string*	ID do *tweet* que iniciou a interação
ironia_final	*string*	Classificação de detecção de ironia
posicao_final	*string*	Classificação de identificação de posicionamento
sentimento_final	*string*	Classificação de análise de sentimento

Para a disponibilização pública[1], o conjunto de dados foi desidratado, ou seja, a base passou a conter apenas os identificadores dos *tweets* e suas respectivas anotações, sem incluir o conteúdo textual original. Essa adaptação foi realizada em conformidade com as diretrizes do X, antigo Twitter, assegurando o respeito à privacidade dos usuários e o cumprimento das políticas de uso da plataforma [Silva *et al.* 2024]. Dessa forma, os textos podem ser recuperados posteriormente por meio da API oficial, preservando tanto a integridade ética quanto a utilidade científica do conjunto de dados.

6. Aplicações

A base de dados possui uma aplicação direta para o projeto HEIWA dentro de sua estratégia de curadoria. É também aplicável em pesquisas voltadas à MA em conteúdos de redes sociais em português brasileiro. A base de dados tem também aplicações para pesquisas com o interesse em analisar dados deste período da história brasileira, bem como para pesquisas que façam uso de dados de redes sociais em português brasileiro.

7. Considerações Finais

Este trabalho se propôs a desenvolver um *dataset* composto por *tweets* no idioma português brasileiro e publicá-lo para utilização no treinamento de modelos de aprendizagem de máquina focados em MA, bem como documentar o processo de coleta e anotação dos dados e a análise das anotações e da base de dados resultante. Devido aos desafios das atividades envolvendo linguagem natural e à natureza do evento aos quais as postagens estão relacionadas, a invasão à sede dos Três Poderes, a base de dados se demonstrou desbalanceada, o que pode limitar sua utilização.

Decorrentes deste trabalho, podem ser exploradas abordagens de anotação com foco na iteratividade, de modo a avaliar o impacto da iteração e refinamento do processo de anotação nas métricas de concordância entre anotadores, além do estudo de técnicas de treinamento de modelos em cenários de bases desbalanceadas, como a utilização do SMOTE para balanceamento de amostras [Suguna *et al.* 2025] e de *Weighted Loss Functions* para adequar o aprendizado do modelo a partir dos dados da base [Cui 2026].

[1] O *dataset* está disponível em zenodo.org/records/17555457, junto ao manual de anotação.

Referências

Artstein, R. e Poesio, M. (2008) "Inter-Coder Agreement for Computational Linguistics". Computational Linguistics, v. 34, n. 4, p. 555–596.

Bertanha, M. C. e Araujo, R. M. (2024) "Linha do Tempo Eleições Presidenciais 2022: evolução da rede social Twitter durante o 8 de janeiro de 2023". Em *Trilha de Temas, Ideias e Resultados Emergentes em Sistemas de Informação – Simpósio Brasileiro de Sistemas de Informação*, p. 339-344. Juiz de Fora/MG. SBC.

Bosc, T.; Cabrio, E. e Villata, S. (2016) "DART: a Dataset of Arguments and their Relations on Twitter". Em *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, p. 1258–1263. European Language Resources Association (ELRA).

Cui, X. (2026) "Addressing Data Imbalance in Transformer-Based Multi-Label Emotion Detection with Weighted Loss". arXiv preprint arXiv:2507.11384.

Fuchs, S. *et al.* (2026) "Human vs. Automated data annotation: Labeling the data set for an ML-driven support ticket classifier". Data & Knowledge Engineering, 162.

Hovy, E. e Lavid, J. (2010) "Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics". International Journal of Translation, v. 22, n.1, p. 13–36.

Krippendorff, K. (2019) Content Analysis: An Introduction to Its Methodology. Sage Publishing.

Lima Filho, S. *et al*. (2024) "DepressSet: Um conjunto de dados de análises textuais sobre postagens depressivas". Em *Brazilian Workshop On Social Network Analysis And Mining* (BRASNAM), 13., p. 214–220. SBC.

Nakayama, H. *et al.* (2018) "Doccano: text annotation tool for human".

Pustejovsky, J. e Stubbs, A. (2012) Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications. O'Reilly Media.

Santos, D. K. S. e Berton, L. (2023) "Analysis of Twitter users' sentiments about the first round 2022 presidential election in Brazil". Em *Encontro Nacional de Inteligência Artificial e Computacional* (ENIAC), 20., p. 880–893. SBC.

Santos, L. A. *et al*. (2025) "Anotação de Dados para a Mineração de Argumentos em Conteúdos de Redes Sociais em Português Brasileiro". Em *Brazilian Workshop On Social Network Analysis And Mining* (BRASNAM), 14., p. 65–78. SBC.

Silva, L. J. *et al*. (2024) "Tweet_Eleições_2022: Um dataset de tweets durante as eleições presidenciais brasileiras de 2022". Em *Brazilian Workshop On Social Network Analysis And Mining* (BRASNAM), 13., p. 193–199. SBC.

Stede, M. e Schneider, J. (2022) Argumentation Mining. Springer Nature.

Suguna, R. *et al*. (2025) "Mitigating class imbalance in churn prediction with ensemble methods and SMOTE". Scientific Reports, 15, n. 16256.

Warrens, M. J. (2010) "Inequalities between multi-rater kappa". Advances in Data Analysis and Classification, v. 4, p. 271–286.