An annotated corpus and computational models for offensive language classification in tweets about the 2022 presidential elections.

Leo Marques Sabino de Freitas[1], **Eloize Rossi Marques Seno**[1], **Sylvia Iasulaitis**[2]

[1]Área de Computação - Instituto Federal de São Paulo – São Carlos, SP

[2]Departamento de Ciências Sociais da Universidade Federal de São Carlos – São Carlos, SP

{leo.freitas@aluno.ifsp.edu.br,eloize@ifsp.edu.br,si@ufscar.br}

Abstract. *Social media has proven to be a fertile environment for the spread of offensive language, especially during elections. Analyzing such interactions is essential to understand behavioral patterns and to propose effective countermeasures. In this context, this article describes the construction of a corpus of tweets related to the 2022 Brazilian elections, annotated for offensive language and hate speech, and evaluates different approaches – including traditional and deep ML models, as well as an LLM – in classifying offensive speech. The results show that traditional approaches perform comparably to more advanced models in classifying types of offenses.*

Resumo. *As redes sociais têm se revelado um ambiente fértil para a propagação de linguagem ofensiva, sobretudo em períodos eleitorais. Analisar tais interações é essencial para compreender padrões de comportamento e propor medidas de enfrentamento eficazes. Neste contexto, este artigo descreve a construção de um corpus de tweets relacionados às eleições brasileiras de 2022, com anotações de ofensas e discurso de ódio, e avalia diferentes abordagens – incluindo modelos tradicionais e profundos de AM, bem como um LLM – na classificação de discurso ofensivo. Os resultados mostram que abordagens tradicionais obtêm desempenhos comparáveis aos de modelos mais avançados na classificação de tipos de ofensas.*

1. Introdução

As redes sociais tornaram-se um espaço central do debate político contemporâneo, permitindo a rápida circulação de opiniões e estratégias eleitorais. Ao mesmo tempo, essa ampliação da comunicação veio acompanhada de um aumento expressivo de comentários ofensivos e discursos de ódio. Marcados por racismo, xenofobia e outras formas de violência verbal, esses discursos degradam a esfera pública e aprofundam a polarização política.

O discurso de ódio é qualquer ato de comunicação que despreze uma pessoa ou grupo por características como raça, cor, etnia, gênero ou orientação sexual [Nockleby 2000]. Embora seja sempre ofensivo, nem todo comentário ofensivo se enquadra como discurso de ódio [Vargas et al. 2022].

Apesar da incitação ao ódio ser considerada crime no Brasil, seu enfrentamento nas mídias sociais ainda é pouco eficaz. Investigar a propagação de linguagem ofensiva

online é, portanto, fundamental para compreender como dinâmicas de interação, poder e conflito se manifestam e se transformam nos ambientes digitais. Esse tipo de análise ajuda a identificar padrões de comportamento e de viralização nas redes, além de fornecer subsídios empíricos e teóricos para o desenvolvimento de estratégias de mitigação dessas práticas, que corroem o debate público e ameaçam a integridade das instituições democráticas. Dado esse contexto, este estudo avaliou diferentes abordagens computacionais para a classificação de linguagem ofensiva e discurso de ódio em *tweets* relacionados ao segundo turno das eleições presidenciais brasileiras de 2022, com vistas à anotação de um grande *corpus* que possa apoiar outros estudos sobre fenômenos sociopolíticos ligados às eleições no país. Foram consideradas três frentes principais: modelos tradicionais de aprendizado de máquina (AM), um modelo de *deep learning* do tipo BERT [Devlin et al. 2019], o BERTimbau [Souza et al. 2020], e o LLM (*Large Language Model*, no inglês) Sabiá-3[1]. Nos modelos tradicionais, adotou-se uma abordagem baseada em AutoML e em atributos comumente utilizados em análise de sentimentos (AS).

Como principais contribuições desta pesquisa destacam-se: i) um *corpus* composto por 3400 *tweets* relacionados ao segundo turno das eleições presidenciais de 2022, manualmente anotados; ii) modelos treinados para a classificação de *tweets* ofensivos e para a classificação dos tipos de ofensa, incluindo discurso de ódio; iii) um *corpus* unificado a partir de *corpora* públicos anotados/revisados por humanos com informações de linguagem ofensiva e discurso de ódio para o treinamento de modelos de AM e iv) resultados de experimentos comparando diferentes abordagens de classificação de linguagem ofensiva e discurso de ódio em comentários de redes sociais em português.

2. Trabalhos Relacionados

Embora várias pesquisas relacionadas à detecção de linguagem ofensiva (ou tóxica) em textos em português tenham sido propostas nos últimos anos ([Leite et al. 2020], [Vargas et al. 2021] [Vargas et al. 2022], [Trajano et al. 2023], [Neto et al. 2024]) e alguns *corpora* anotados foram disponibilizados para o treinamento de modelos, ainda há uma carência de estudos e de dados rotulados para domínios específicos como o político.

O maior *corpus* com anotações manuais de linguagem ofensiva, do qual se tem conhecimento, é o TOLD-Br [Leite et al. 2020], composto de 21 mil *tweets* coletados em 2019 e anotados em sete categorias. As anotações foram realizadas por 48 colaboradores, com três avaliadores por *tweet*, apresentando concordância média de 55% (Krippendorff's Alpha). O corpus foi utilizado para treinar modelos BERT na classificação binária (conteúdo tóxico versus não tóxico). O BERTimbau obteve o melhor desempenho dentre os modelos investigados, alcançando macro-F1 de 76%.

No trabalho de [Vargas et al. 2021] a detecção de linguagem abusiva nas redes sociais é abordada em duas tarefas distintas: i) identificação de comentários ofensivos e ii) detecção de discurso de ódio. Os autores treinaram dois classificadores, um para cada tarefa. Entre as *features* usadas destaca-se o MOL - *Multilingual Offensive Lexicon*, um léxico de palavras e expressões ofensivas explícitas e implícitas em português. Para os experimentos, os autores construíram o *corpus* Hate-Br [Vargas et al. 2022], composto por 7 mil comentários coletados do Instagram direcionados a políticos brasileiros.

[1] `https://www.maritaca.ai/`

O *corpus* foi anotado manualmente em três níveis, a saber: classificação binária de comentários ofensivos, classificação do nível de ofensa (leve, moderada ou alta), e nove classes de discurso de ódio . Os autores relataram 88% de macro F-*score* na classificação binária usando o modelo Naive Bayes, e 85% de F-*score* na classificação de discurso de ódio (multi-rótulos), usando o classificador MLP. De forma similar ao trabalho de [Vargas et al. 2021], neste estudo a detecção de linguagem ofensiva e discurso de ódio é abordada em duas tarefas distintas, a saber: classificação de *tweets* ofensivos e não ofensivos e classificação dos tipos de ofensas, incluindo categorias de ódio (vide seção 3). Outra relação com esse trabalho, se refere ao uso do MOL como *feature* no treinamento de modelos com AutoML (vide seção 4.2).

Outro trabalho que tem relação com esta pesquisa é o de [Neto et al. 2024]. Os autores propuseram o *corpus* Toxic-BR com 143.854 *tweets* relacionados ao período eleitoral das eleições presidenciais de 2022, construído usando uma abordagem semi-automática. Cada *tweet* foi classificado de forma binária em tóxico e não tóxico por dois LLMs: a Perspective API da Jigsaw/Google e uma versão ajustada do ChatGPT 3.0[2] usando o TOLD-Br. Posteriormente, um total de 1400 amostras foram aleatoriamente selecionadas para revisão humana, sendo que destas 1 mil tinham rótulos divergentes atribuídos pelos modelos, 200 *tweets* foram classificados por ambos modelos como tóxicos e outros 200 foram classificados pelos dois como não tóxicos. O *corpus* revisado contém um total de 615 *tweets* tóxicos e 785 *tweets* não tóxicos. De forma similar, esta pesquisa também investigou o uso de um LLM - na classificação de *tweets* sobre as eleições de 2022, porém optou-se aqui por usar um modelo monolíngue treinado especificamente com textos em português (vide seção 4.3).

3. Construção do Corpus

A construção do *corpus* do segundo turno das eleições presidenciais de 2022 se deu a partir da extração de *tweets* do *dataset* ITED-BR [Iasulaitis et al. 2025], composto por 228 milhões de mensagens coletadas da rede social X (na época Twitter) durante o período eleitoral. A extração se baseou na coleta de *tweets* contendo termos e *hashtags* relacionados aos dois candidatos ao segundo turno, Jair Messias Bolsonaro (daqui em diante, Bolsonaro) e Luiz Inácio Lula da Silva (daqui em diante, Lula), publicados entre 02 e 30 de outubro de 2022, período que sucede e antecede o primeiro e o segundo turnos, respectivamente. Ao todo foram coletadas 11.175.094 de mensagens dirigidas ao Bolsonaro e 11.501.709 de mensagens relacionadas ao Lula. A coleta se baseou no uso dos seguintes termos relativos a cada candidato:

> **Bolsonaro:** *bolsonaro, jair bolsonaro, bozo, biroliro, tchutchuca do centrão, bonoro, capitão, genocida, mito, bolsomito, bolsolixo, bolsotrump, messias, patriota, b22, b17, brocha, imbrochavel, maçonaro.*
> **Lula:** *lula, ex-presidiário, lulalivre, 9 dedos, luladrão, lulaladrão, lulinha, nine, luis inácio, cachaceiro, sapo barbudo, lulão, l13, faz o L, lulindo, metalúrgico, lulalkimin.*

Devido à limitação de recursos computacionais para o processamento do enorme volume de dados coletados, foram selecionados aleatoriamente 50k *tweets* relativos a cada candidato. Para evitar a coleta de comentários pouco informativos, foram selecionados *tweets*

[2] `https://chatgpt.com/`

com no mínimo 10 *tokens*, usando o tokenizador do SpaCy[3].

Com o intuito de construir um *corpus* de referência para a avaliação das abordagens e modelos investigados neste estudo, foram selecionados aleatoriamente 3400 *tweets* para anotação manual (1700 de cada candidato). A anotação foi realizada por nove colaboradores do projeto FAPESP (nº 2022/03090-0). Para mitigar vieses subjetivos, a equipe abrangeu diversidade demográfica (4 homens, 5 mulheres; 6 brancos, 3 pardos; diferentes orientações sexuais e formações acadêmicas e pluralidade política (da extrema esquerda à direita). A anotação baseou-se nas diretrizes propostas por [Trajano et al. 2023]. Durante o processo, cada *tweet* foi inicialmente classificado como ofensivo (1) ou não ofensivo (0). Em seguida, o *tweet* ofensivo foi classificado em uma ou mais categorias, de acordo com o tipo de ofensa apresentado. Sete categorias foram usadas: *insult*, *obscene*, *ideology*, *racism*, *sexism*, *xenophobia* e *LGBTQphobia*. Além dessas sete, os anotadores eram instruídos a classificar como "*outra*", caso identificassem outras categorias. Esse conjunto resulta da síntese das principais classes consolidadas na literatura com maior propensão de ocorrência no domínio considerado.

Os dois conjuntos de 1700 *tweets* foram divididos em subconjuntos menores, contendo 100 *tweets* cada. Posteriormente, cada subconjunto foi atribuído a um grupo composto por três anotadores. Para viabilizar o cálculo da concordância, 25% dos *tweets* de cada anotador eram iguais aos dos demais anotadores do grupo, enquanto que os 75% restantes eram anotados por somente uma pessoa. A classe final dos *tweets* comuns aos três representa o voto da maioria. Nos casos em que os três anotadores discordavam em relação aos tipos de ofensas, as divergências eram discutidas e, persistindo o desempate, a classe final era definida por um quarto anotador.

A fim de se obter uma anotação mais homogênea, os integrantes de cada grupo se alternavam a cada novo subconjunto a ser analisado. Antes de iniciar a anotação, foi feito um breve treinamento com um conjunto de 20 *tweets* (10 de cada candidato) que foram anotados por todos os anotadores conjuntamente. A concordância média geral, calculada com o coeficiente Krippendorff's Alpha, foi de 69,0% para os *tweets* dirigidos ao Lula e 61,0% para os *tweets* relativos ao Bolsonaro, indicando um alto nível de confiabilidade na anotação.

A Tabela 1 apresenta a distribuição de exemplos por categoria anotados no *corpus* para cada conjunto de *tweets* e a distribuição total. Observa-se uma predominância de exemplos ofensivos (i.e. 68,3% no *corpus* todo). Em relação aos tipos de ofensas, nota-se uma prevalência de comentários contendo insultos, ofensas e palavrões (i.e. 95,8%) e de discursos de ódio baseados em ideologias, especialmente, político-partidária (18,8%). Com relação à categoria *outras* (38 ocorrências em todo *corpus*), 14 ocorrências (36,8%) foram classificadas como intolerância religiosa, 11 (28,9%) como capacitismo, 7 (18,4%) como etarismo e 6 (15,9%) como aporofobia. A Tabela 2 apresenta exemplos de *tweets* do *corpus* anotados manualmente.

Para os experimentos realizados neste estudo (vide seção 5), os dois conjuntos de *tweets* foram concatenados em um único *corpus* e pré-processados para a remoção de URLs, *hashtags*, marcas de *retweet* (RT) e substituição de menções a usuários por `@user`. A normalização ortográfica, necessária para lidar com gírias, abreviações e variações de

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

Tabela 1. Total de *tweets* por classe em cada conjunto do *corpus*.

Ofensa	Bolsonaro	Lula	Total
Insult	1188 (95,4%)	1036 (96,2%)	2224 (95,8%)
Ideology	244 (20,5%)	193 (18,0%)	437 (18,8%)
Obscene	106 (8,9%)	70 (6,5%)	176 (7,6%)
Other	24 (1,4%)	14 (0,8%)	38 (1,1%)
Xenophobia	12 (1,1%)	5 (0,4%)	17 (0,7%)
Sexism	11 (0,9%)	5 (0,4%)	16 (0,7%)
LGBTQphobia	8 (0,7%)	0 (0,0%)	8 (0,3%)
Racism	3 (0,2%)	1 (0,09%)	4 (0,2%)
#Ofensivos	1245 (73,2%)	1076 (63,3%)	2321 (68,3%)
#Não ofensivos	455 (26,8%)	624 (36,7%)	1079 (31,7%)

escrita foi realizada com a ferramenta Enelvo [Bertaglia and Nunes 2016].

Tabela 2. Exemplos de *tweets* ofensivos extraídos do *corpus*.

Tweet	**Conjunto**	**Classes**
"Mais o Lula é ex-presidiário" o teu tbm é putinha	Lula	*insult*, *obscene*, *sexism*
@user @user @user A cloaca senil e o cachaceiro se merecem. Mas o Brasil não merece esse escárnio. A máfia do ladrão ésta desesperada, kkkkkkk	Lula	*insult*, *obscene* *ideology*
@user @user É isso, pra vcs verem que beleza não representa inteligência kkkkk Uma mulher linda, que diz que estuda e vota no Bolsonaro mesmo assim. Certeza que as notas de história eram vermelhas	Bolsonaro	*insult* *ideology*, *sexism*
Por isso eu digo: pior que o Bolsonaro, só mesmo quem ainda apoia ele... E outra coisa: pobres, gays, mulheres, negros e quem perdeu alguém pra covid, mas que seguem apoiando esse cara, são as pessoas mais burras que existem!	Bolsonaro	*insult*, *LGBTQphobia*, *racism*, *sexism*

4. Classificação automática de linguagem ofensiva

As duas tarefas de classificação abordadas neste estudo foram avaliadas usando um modelo BERT, vários modelos treinados com AutoML e um LLM generativo. O treinamento com o BERT e o AutoML se baseou no uso de *corpora* de referências disponíveis na literatura (vide seções 4.1 e 4.2). Os modelos treinados foram avaliados usando o *corpus* descrito na seção 3. A Figura 1 ilustra o fluxo geral de treinamento e avaliação dos modelos supervisionados. As subseções a seguir descrevem como cada abordagem foi usada em cada tarefa.

4.1. Treinamento de modelos BERT

O BERT (*Bidirectional Encoder Representations from Transformers*) é um modelo de *deep learning* baseado em uma abordagem de *transfer learning* que permite o ajuste fino (*fine-tuning*) do modelo pré-treinado em grandes *corpora* para tarefas específicas com menos dados rotulados. Nesta pesquisa foi usado o BERTimbau Large[4], pré-treinado para o português, para treinar modelos específicos para a classificação binária de comentários ofensivos e não ofensivos e para a classificação dos tipos de ofensa. A tarefa binária é uma etapa preliminar para a categorização de ofensas.

Para o *fine-tuning* da tarefa binária foi utilizado o HateBr [Vargas et al. 2022] (vide seção 2). A escolha desse *corpus* se deve ao fato dele ser balanceado (com 3500 comentários ofensivos e 3500 comentários não ofensivos) e do domínio político. O *fine-tuning* foi realizado em ambiente com GPU, usando 65% do *corpus* na etapa de treinamento, 20% na validação e 15% no teste. Os seguintes hiperparâmetros foram adotados:

[4]`https://huggingface.co/neuralmind/bert-large-portuguese-cased`

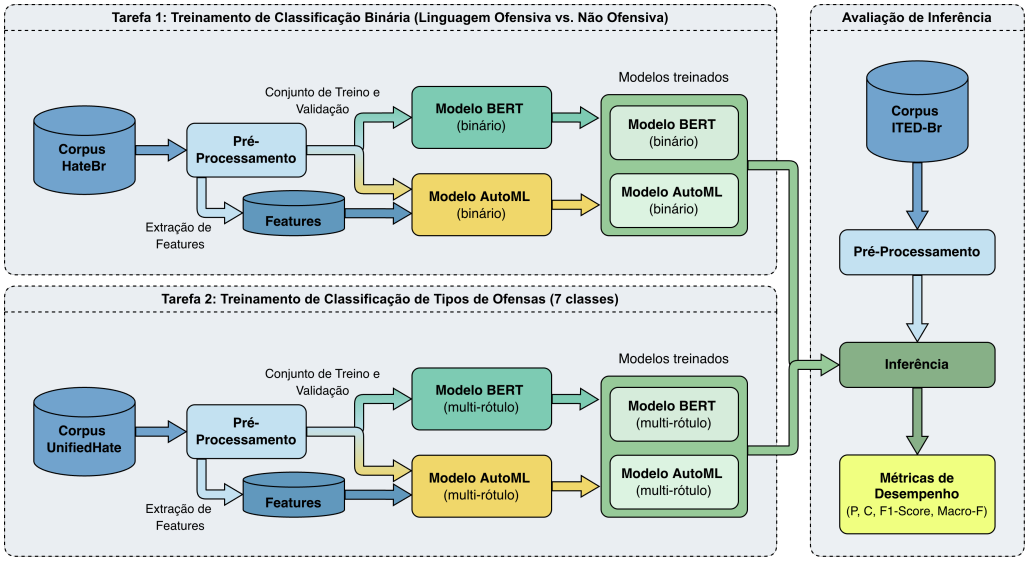


Figura 1. Fluxo geral das tarefas de classificação supervisionadas.

batch size de 32, 3 épocas, *weight_decay* de 0.01, *learning_rate* = 2e-5, *warmup_ratio* = 0.1, *max_sequence_length* = 128 e *label_smoothing_factor* = 0.1, *early_stopping_patience* = 2 e *early_stopping_threshold* = 0.0, com avaliação ao final de cada época. O HateBr passou pelo mesmo pré-processamento do *corpus* do segundo turno (vide seção 3). No conjunto de teste, o modelo alcançou 92,0% de *F1-score* em ambas as classes (92,0% de macro-F), evidenciando, assim, sua capacidade de discriminar comentários ofensivos e não ofensivos.

Para o *fine-tuning* na tarefa de classificação dos tipos de ofensas, foram usados três *corpora* públicos consolidados na literatura, a saber: HateBr [Vargas et al. 2022]; TOLD-Br [Leite et al. 2020], composto de 21 mil *tweets*, sendo cerca de 44% deles ofensivos; e OLID-Br [Trajano et al. 2023], composto de 6.354 comentários coletados de diferentes plataformas (Twitter, YouTube, etc.) com anotações de linguagem ofensiva e suas categorias. A fim de se obter um *corpus* mais representativo para o treinamento do modelo nessa tarefa, os três *corpora* foram unificados e os comentários de categorias semanticamente relacionadas, mas com diferentes rótulos nos três *datasets* (como *homophobia* e *LGBTQphobia*, *partyism* e *ideology*) foram agrupados em uma mesma categoria mais genérica. Na unificação, foram consideradas apenas as classes comuns a dois ou mais *corpora* e que apresentavam distribuição significativa no *corpus* de origem. As mensagens repetidas foram descartadas. O *corpus* resultante, aqui nomeado de *UnifiedHate*, contém um total de 15.555 comentários. A Tabela 3 apresenta cada classe e a sua distribuição no *UnifiedHate*, bem como a classe original e o total de ocorrências no *corpus* de origem.

Tabela 3. Distribuição de classes no *corpus* UnifiedHate.

Classe	Told-BR	Olid-BR	Hate-BR	Total	Classe Original
Insult	4.371	5.589	-	9.864 (63.41%)	Insult, Insult
Obscene	6.639	2.311	-	8.819 (56.70%)	Obscene, Profanity/Obscene
Ideology	-	1.221	496	1.704 (10.95%)	Ideology, Partyism
LGBTQphobia	344	375	17	731 (4.70%)	LGBTQ+phobia, LGBTQphobia, Homophobia
Racism	138	92	8	236 (1.52%)	Racism, Racism, Racism
Sexism	463	461	97	1.011 (6.50%)	Misogyny, Sexism, Sexism
Xenophobia	151	128	1	279 (1.79%)	Xenophobia, Xenofobia, Xenophobia

Para o treinamento, foram usados os mesmos hiperparâmetros adotados na classificação binária e a mesma divisão de treinamento, validação e teste. Além disso, o *UnifiedHate* passou pelo mesmo pré-processamento mencionado na seção 3. A Tabela 4 apresenta os valores de macro precisão (Macro-P), macro cobertura (Macro-C) e macro F1-*score* (macro-F1) obtidos pelo BERTimbau na classificação dos tipos de ofensas no conjunto de teste. Para a maioria das classes, o modelo alcançou valores de macro-F1 entre 75,0% e 84,0%, tendo apresentado um pior desempenho na classificação de comentários racistas (50,0%) e xenofóbicos (54,0%), o que já se esperava por se tratarem de classes minoritárias. O desempenho global do modelo (Macro-F), considerando a média dos F-*scores* de todas as classes, foi de 73,4%.

Tabela 4. Resultados obtidos pelo BERT na classificação dos tipos de ofensas.

	Insult	**Obscene**	**Ideology**	**LGBTQphobia**	**Racism**	**Sexism**	**Xenophobia**
Macro-P	0.83	0.84	0.92	0.87	0.49	0.83	0.99
Macro-C	0.85	0.84	0.78	0.81	0.50	0.70	0.52
Macro-F1	**0.84**	**0.84**	**0.83**	**0.84**	0.50	**0.75**	0.54
Suporte	1480	1317	198	98	35	144	42

4.2. Treinamento de modelos com AutoML

Para o treinamento com o AutoML (*Automated Machine Learning*) foi usada a biblioteca *open-source* AutoGluon[5], que automatiza o treinamento de modelos de aprendizado de máquina sem que haja necessidade de intervenção manual na seleção de algoritmos e na configuração de hiperparâmetros. Dentre as *features* investigadas estão o MOL conforme proposto por [Vargas et al. 2021], que utiliza o léxico de ofensas do português, o MOL, e vários atributos baseados em técnicas de análise de sentimentos (AS). Mais especificamente, as *features* de AS se baseiam no GoEmotions [Hammes and Freitas 2021], um BERT pré-treinado para a análise de emoções em textos, e no uso de léxicos de sentimentos, que dispõem de conjuntos de palavras associadas com polaridade (positiva ou negativa) e emoções (como raiva, alegria e decepção). A motivação para o uso dessas técnicas é a de que comentários ofensivos tendem a expressar sentimentos e emoções negativas.

O GoEmotions recebe um texto de entrada (ex. *tweet*) e produz como saída uma lista de emoções e suas probabilidades de ocorrência no texto. Ele detecta até 27 emoções mais a emoção neutra. Duas *features* distintas foram obtidas a partir desse modelo. A primeira é baseada nos *scores* de probabilidade das emoções previstas pelo modelo, representados em um vetor de probabilidades $E = \{p_1, p_2, ..., p_{28}\}$, onde p_i representa a probabilidade associada à emoção i (daqui em diante, *feature* GE). A segunda *feature* é baseada nas emoções com maior probabilidade de ocorrência, considerando um limiar de corte de 0.3 definido empiricamente. Cada emoção foi mapeada para uma polaridade positiva (1), negativa (-1) ou neutra (0), seguindo o mapeamento proposto por [Assi et al. 2022]. A polaridade final de um texto é dada pela soma das suas polaridades positivas e negativas (daqui em diante, *feature* GEP).

Por fim, foram definidas outras duas *features* baseadas em três léxicos de sentimentos de referência na literatura, a saber: LIWC-Br [Filho Balage et al. 2013] , Word-

[5]https://github.com/autogluon/autogluon

NetAffectBR [Pasqualotti and Vieira 2008] e SentiLex-PT [Silva et al. 2012]. Inicialmente, foram identificadas todas as palavras de cada comentário presentes em pelo menos um dos léxicos e suas respectivas polaridades. Em seguida, foram calculadas as proporções de termos negativos, positivos e neutros de cada comentário. Essas proporções correspondem à razão entre o número de *tokens* associados a cada polaridade e o número total de *tokens* do texto (daqui em diante, *feature* Lex). Esse cálculo foi realizado após a remoção de *stopwords*. A última *feature* é um índice composto dessas três proporções de polaridade (índice IP), calculado com base na Equação 1 (daqui em diante, *feature* IP).

$$IP = \beta_0 + \beta_1 \cdot feat_{neg} + \beta_2 \cdot feat_{pos} + \beta_3 \cdot feat_{neu} \quad (1)$$

onde $feat_{neg}$, $feat_{pos}$ e $feat_{neu}$ representam, respectivamente, as proporções de termos negativos, positivos e neutros. Os coeficientes β foram estimados usando regressão logística, onde β_0 corresponde ao intercepto (ou *bias*) do modelo, β_1, β_2 e β_3 correspondem aos coeficientes associados às *features* $feat_{neg}$, $feat_{pos}$ e $feat_{neu}$, respectivamente. Quanto às variáveis-alvo adotadas nas modelagens, para os modelos binários foi utilizada a classe *ofensivo* como variável-alvo para a estimação dos coeficientes. Já na classificação de tipos de ofensas, foi utilizada a classe majoritária (*insulto*).

Tabela 5. Resultados obtidos com AutoML na tarefa binária com *HateBr*.

Modelo	Não ofensivo			Ofensivo			
	P	C	F1	P	C	F1	Macro-F
Baseline	**0.85**	**0.91**	**0.88**	**0.90**	**0.83**	**0.86**	**0.87**
MOL+GE+GEP+Lex+IP	**0.85**	0.88	**0.87**	0.87	**0.84**	**0.86**	**0.86**
MOL+GE	**0.84**	0.88	0.86	0.86	**0.83**	**0.85**	0.85
MOL+GE+GEP+Lex	**0.85**	0.88	0.86	0.87	**0.84**	**0.85**	0.85
MOL+GE+GEP	**0.84**	0.87	0.86	0.86	**0.83**	**0.85**	0.85
MOL	0.76	0.81	0.79	0.80	0.74	0.77	0.78

Para o treinamento de modelos de classificação binária e de classificação de tipos de ofensas foram usados os *corpora* Hate-Br e o *UnifiedHate*, respectivamente. Com o objetivo de compreender a contribuição de cada *feature*, foi feita uma avaliação incremental iniciando com um modelo básico do AutoML sem atributos (*Baseline*) e incluindo um atributo por vez no modelo, o que resultou em seis diferentes configurações, a saber: *Baseline*, MOL, MOL+GE, MOL+GE+GEP, MOL+GE+GEP+Lex e MOL+GE+GEP+Lex+IP. Os resultados obtidos na classificação binária e de tipos de ofensas nos conjuntos de testes são apresentados nas Tabelas 5 e 6, respectivamente. De modo geral, há pouca variação de desempenho em termos de Macro-F, quando se compara as diferentes configurações de modelos, nas duas tarefas de classificação. Há exceção apenas na classificação binária em que se percebe uma queda de desempenho de 11 pontos percentuais ao incluir o MOL como atributo único (i.e. 87,0% de Macro-F obtido pelo *Baseline* contra 78,0% obtido pelo MOL). Apesar da pequena diferença de desempenho entre as configurações do AutoML, nota-se (Tabela 5) que o *Baseline* alcançou a maior cobertura na classe *não ofensivo* (91,0%) e a maior precisão na classe *ofensivo* (90,0%). Em relação à segunda tarefa (Tabela 6), como esperado, os modelos tiveram um desempenho pior nas classes minoritárias, ou seja, *racismo* e *xenofobia*, enquanto que para as demais classes os valores de F1 ficaram acima de 65,0%, ultrapassando os 80,0% na classificação de *insultos* e comentários *obscenos* (vulgares), valores comparáveis aos obtidos pelo BERT (vide Tabela 4).

Tabela 6. Resultados obtidos com AutoML na classificação de tipos de ofensas.

Modelo	Insult			Obscene			Ideology			LGBTQphobia			Racism			Sexism			Xenophobia			M-F
	P	C	F1	P	C	F1	P	C	F1	P	C	F1	P	C	F1	P	C	F1	P	C	F1	
Baseline	0.82	0.80	**0.81**	0.83	0.83	**0.83**	0.81	0.77	**0.79**	0.83	0.64	**0.69**	0.55	0.52	0.53	0.71	0.64	**0.66**	0.57	0.52	0.53	**0.69**
MOL	0.82	0.81	**0.81**	0.83	0.83	**0.83**	0.81	0.78	**0.79**	0.83	0.66	**0.71**	0.49	0.50	0.50	0.66	0.66	**0.66**	0.49	0.50	0.50	**0.69**
MOL+GE+GEP+Lex	0.83	0.82	**0.82**	0.82	0.83	**0.82**	0.76	0.82	**0.78**	0.76	0.70	**0.72**	0.49	0.49	0.49	0.72	0.65	**0.68**	0.49	0.50	0.49	**0.69**
MOL+GE+GEP+Lex+IP	0.83	0.82	**0.83**	0.82	0.82	**0.82**	0.77	0.81	**0.79**	0.69	0.70	**0.70**	0.52	0.51	0.51	0.68	0.67	**0.68**	0.74	0.52	0.54	**0.69**
MOL+GE	0.83	0.82	**0.82**	0.83	0.83	**0.83**	0.77	0.76	**0.76**	0.78	0.68	**0.72**	0.49	0.50	0.50	0.67	0.67	**0.67**	0.49	0.50	0.49	**0.68**
MOL+GE+GEP	0.83	0.82	**0.82**	0.82	0.83	**0.82**	0.76	0.79	**0.77**	0.81	0.66	**0.71**	0.49	0.50	0.50	0.72	0.64	**0.66**	0.68	0.53	0.56	**0.68**

4.3. LLM Sabiá

Outra abordagem investigada baseou-se no uso do Sabiá-3[6], um LLM monolíngue treinado com textos em português. Trata-se de um modelo generativo baseado em *prompt* que recebe como entrada um texto contendo a descrição da tarefa a ser realizada e gera as saídas conforme solicitado. O maior desafio em lidar com esse tipo de modelo consiste em definir um *prompt* que gere as saídas esperadas. Depois de várias tentativas, incluindo *prompts zero-shot* e *few-shot*, percebeu-se que em ambas as tarefas de classificação o modelo desempenhava melhor com *few-shot*. Os *prompts* usados na classificação binária e na classificação de tipos de ofensas são apresentados a seguir, respectivamente:

Prompt 1 (Classificação Binária): *Você é um classificador de tweets ofensivos. Dado um tweet, responda apenas com 1 se for ofensivo ou com 0 se não for ofensivo. Aqui estão 2 exemplos de tweets e suas classificações (entre parênteses): 1. "Minas Gerais é o estado com maior número de denúncias de crime e assédio eleitoral. A maioria esmagadora das denúncias tem conexão com a campanha de Bolsonaro, com a anuência do governador, que também é acusado de fazer assédio a prefeitos e políticos mineiros." (0) 2. "@user @user Não adianta derrubar! Todo mundo já sabe: Lula é ladrão. homicida, corrupto, mentiroso, ateu, terrorista, chefe de organizações criminosas!" (1)*

Prompt 2 (Tipos de Ofensas): *Você é um classificador de tweets ofensivos. Dado um tweet, responda com os seguintes rótulos (0 para ausência, 1 para presença): insult: insulto, palavrões ou ofensas cujo objetivo é provocar, humilhar ou atacar alguém verbalmente; obscene: conteúdo vulgar, sexualmente explícito ou pornográfico; ideology: discurso de ódio baseado em ideologias; lgbtqphobia: discurso de ódio contra identidade de gênero ou orientação sexual; racism: discurso de ódio baseado em raça, etnia ou cor; sexism: discurso de ódio baseado em gênero ou sexo; xenophobia: discurso de ódio contra estrangeiros ou pessoas de outras culturas ou nacionalidades. Responda no formato: insult:0,obscene:0,ideology:0,lgbtqphobia:0,racism:0,sexism:0,xenophobia:0*

5. Resultados dos experimentos e discussão

O *corpus* de *tweets* relacionado ao segundo turno das eleições de 2022 foi usado na avaliação do Sabiá-3 e dos modelos de inferência obtidos com o BERT e o AutoML. Os resultados da classificação binária são apresentados na Tabela 7. O melhor desempenho foi obtido pelo BERT, enquanto que o segundo melhor desempenho foi obtido com o LLM Sabiá. Contudo, os modelos de AM tradicionais baseados no MOL associado com *features* de análise de sentimentos ficaram atrás do LLM e do BERT por apenas 4 e 6 pontos percentuais, respectivamente, alcançando 69,0% de macro-F. Ao comparar as diferentes configurações do AutoML, nota-se que o GoEmotions (atributo GE) contribuiu para um ganho de desempenho em relação ao *Baseline* de 6 pontos percentuais (69,0% de macro-F contra 63,0%) e de 7 pontos percentuais comparado ao MOL (69,0% de macro-F contra 62,0%), revelando a importância desse atributo para a discriminação de classes do modelo. Por outro lado, observa-se que as *features* GEP, Lex e IP tiveram pouca influência no desempenho dos modelos.

A Tabela 8 apresenta os resultados obtidos na classificação de tipos de ofensas em termos de macro precisão (P), macro cobertura (C), F1-*score* (F1) e macro-F (M-F). O melhor macro-F foi obtido pelo Sabiá (58,0%), seguido pelo *baseline* (55,0%). Contudo,

[6] https://www.maritaca.ai/

essa abordagem simples do AutoML superou o LLM na classificação de comentários obscenos (76,0% de F1 contra 65,0%) e obteve desempenho comparável ao do LLM nas demais classes, com exceção da classe *insult*. O BERT só superou esses modelos na classe *insult* (80,0% de F1), apresentando o pior Macro-F (28,0%). O desempenho inferior do BERT pode ser atribuído ao acentuado desbalanceamento de classes no *corpus*. Além disso, não se descarta a hipótese de um leve *overfitting* aos dados de treino, ainda que a técnica de *early stopping* tenha sido empregada na tentativa de mitigar esse comportamento (vide seção 4.1).

Ao incluir os atributos MOL, GE, GEP e Lex no AutoML, houve uma melhora substancial na precisão da classe *insult*, comparado ao *Baseline*, saltando de 59,0% para 68,0%, e também em termos de F1, saltando de 41,0% para 45,0%. Houve, ainda, melhora na classe *sexism*, tanto na precisão como na cobertura, levando a um aumento no valor de F1 de 56,0% para 61,0%. Contudo, essas *features* levaram a uma queda no desempenho das classes *obscene* e *LGBTQphobia*.

Para a classificação binária dos *tweets* restantes do *corpus* (vide seção 3) foi usado o BERTimbau e para a classificação de tipos de ofensas foi usado o *Baseline*, que apresentou resultados competitivos com os do LLM[7].

Tabela 7. Resultados obtidos na classificação binária.

Modelo	Não ofensivo			Ofensivo			
	P	C	F1	P	C	F1	Macro-F
BERT	**0.64**	**0.70**	**0.67**	**0.85**	**0.81**	**0.83**	**0.75**
Sabiá	**0.59**	**0.71**	**0.65**	**0.85**	**0.77**	**0.81**	**0.73**
MOL+GE	0.54	0.67	**0.60**	**0.83**	**0.74**	**0.78**	**0.69**
MOL+GE+GEP+Lex	0.54	0.67	**0.60**	**0.83**	**0.74**	**0.78**	**0.69**
MOL+GE+GEP+Lex+IP	0.55	0.65	**0.60**	**0.82**	**0.75**	**0.78**	**0.69**
MOL+GE+GEP	0.53	0.68	**0.60**	**0.83**	0.72	0.77	0.68
Baseline	0.47	0.62	0.53	0.79	0.67	0.73	0.63
MOL	0.45	0.70	0.55	0.81	0.61	0.70	0.62

Tabela 8. Resultados obtidos na classificação de tipos de ofensas.

Modelo	Insult			Obscene			Ideology			LGBTQphobia			Racism			Sexism			Xenophobia			M-F
	P	C	F1	P	C	F1	P	C	F1	P	C	F1	P	C	F1	P	C	F1	P	C	F1	
Sabiá	0.70	0.72	0.68	**0.80**	0.62	**0.65**	**0.57**	**0.65**	**0.53**	**0.56**	**0.68**	**0.59**	0.51	0.62	0.52	**0.54**	**0.62**	**0.55**	0.54	0.62	0.55	**0.58**
Baseline	0.59	0.51	0.41	**0.73**	**0.80**	**0.76**	**0.50**	**0.50**	**0.49**	**0.67**	**0.57**	**0.59**	0.50	0.50	0.50	**0.55**	**0.59**	**0.56**	0.51	0.52	0.51	**0.55**
MOL	0.58	0.51	0.42	0.73	0.81	0.76	0.50	0.50	0.49	0.58	0.62	0.60	0.50	0.50	0.50	0.54	0.65	0.56	0.50	0.50	0.50	0.55
MOL+GE	0.67	0.52	0.44	0.68	0.80	0.72	0.50	0.50	0.49	0.52	0.62	0.53	0.50	0.50	0.50	0.55	0.71	0.58	0.50	0.50	0.50	0.54
MOL+GE+GEP	0.63	0.52	0.45	0.68	0.80	0.72	0.50	0.50	0.49	0.53	0.56	0.54	0.50	0.50	0.50	0.56	0.59	0.57	0.50	0.50	0.50	0.54
MOL+GE+GEP+Lex	**0.68**	0.52	**0.45**	0.69	0.79	0.72	0.50	0.50	0.49	0.53	0.68	0.54	0.50	0.61	0.50	**0.58**	**0.68**	**0.61**	0.50	0.50	0.50	0.54
MOL+GE+GEP+Lex+IP	0.67	0.52	0.45	0.66	0.80	0.70	0.50	0.49	0.48	0.51	0.67	0.52	0.50	0.49	0.50	0.55	0.68	0.58	0.50	0.50	0.50	0.53
BERT	0.68	**0.96**	**0.80**	0.46	0.64	0.53	0.16	0.10	0.12	0.09	0.12	0.11	0.00	0.00	0.00	0.27	0.19	0.22	0.33	0.12	0.17	0.28

6. Conclusão

Este artigo descreveu a anotação manual de um *corpus* de *tweets* relacionados ao segundo turno das eleições presidenciais de 2022, ocorrida no Brasil, com informações sobre linguagem ofensiva e discurso de ódio e investigou o desempenho de diferentes abordagens computacionais — desde modelos clássicos de aprendizado de máquina (AM) até modelos mais avançados como BERT e LLM — na classificação binária de *tweets* ofensivos e na classificação de tipos de ofensas. Na tarefa binária, os melhores desempenhos foram obtidos com um BERT ajustado para a tarefa e com o LLM Sabiá, ou seja, 75,0% e 73,0% de macro-F, respectivamente. Contudo, os modelos tradicionais de AM, que são

[7]https://github.com/leomsfreitas/HateSpeech

computacionalmente mais simples e baratos do que o BERT e o LLM, apresentaram um desempenho comparável ao desses modelos atingindo 69,0% de macro-F.

Na classificação de tipos de ofensas, mais uma vez os experimentos mostraram a relevância das abordagens tradicionais. Embora o melhor resultado tenha sido obtido pelo Sabiá (58,0% de macro-F), o modelo básico treinado com AutoML se mostrou bastante competitivo com esse modelo, alcançando 55,0% de macro-F, enquanto o BERT obteve o pior desempenho (28,0% de macro-F), tendo sido capaz de generalizar apenas na classificação de insultos (classe majoritária).

Como trabalho futuro, o *corpus* construído nesta pesquisa e os modelos treinados serão incorporados em um *framework* de *Active Learning*, com o intuito de expandir a quantidade de exemplos rotulados do *corpus* e melhorar o desempenho dos modelos nas duas tarefas de classificação.

7. Considerações éticas

Este trabalho trata de dados sensíveis provenientes de mídias sociais relacionados ao contexto da política brasileira, o que exige cuidados éticos rigorosos. Para preservar a privacidade dos usuários, os dados foram anonimizados no pré-processamento (conf. seção 3). O compartilhamento e a redistribuição dos dados respeitam as diretrizes e políticas de uso previstas para o *dataset* ITED-Br [Iasulaitis et al. 2025], garantindo o acesso responsável ao *corpus* para fins estritamente acadêmicos e de pesquisa. Adicionalmente, reconhece-se o impacto da exposição contínua de anotadores humanos a conteúdos potencialmente nocivos e ofensivos. Para mitigar esse desgaste psicológico e viabilizar um ambiente de trabalho seguro, a etapa de anotação contou com treinamento prévio, suporte entre os membros da equipe e distribuições de rotulagem intercaladas em pequenos lotes.

8. Agradecimentos

Nossos agradecimentos à FAPESP pelo apoio ao projeto nº 2022/03090-0, ao IFSP pelo apoio financeiro na forma de bolsa de IC, à Maritaca AI pelos créditos concedidos para uso do LLM e aos anotadores do *corpus* pela valiosa contribuição.

Referências

Assi, F. M., Candido, G. B., Silva, L. N. d. S., Silva, D. F., and Caseli, H. d. M. (2022). Ufscar's team at absapt 2022: Using syntax, semantics and context for solving the tasks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*, volume 3202 of *CEUR Workshop Proceedings*.

Bertaglia, T. F. C. and Nunes, M. d. G. V. (2016). Exploring word embeddings for unsupervised textual user-generated content normalization. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 112–120.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Filho Balage, P. P., Pardo, T. A. S., and Aluísio, S. M. (2013). An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

Hammes, L. O. A. and Freitas, L. A. d. (2021). Utilizando bertimbau para a classificação de emoções em português. In *Anais do Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*.

Iasulaitis, S., Valejo, A. D. B., Greco, B. C., Perillo, V. G., Messias, G. H., Vicari, I., with the Interfaces—Center for Sociopolitical Studies of Algorithms, and Intelligence, A. (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):1–29.

Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In Wong, K.-F., Knight, K., and Wu, H., editors, *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*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Neto, F. R., Anchiêta, R., Moura, R., and Santana, A. (2024). Abordagem semi-supervisionada para anotação de linguagem tóxica. In *Anais do XIII Brazilian Workshop on Social Network Analysis and Mining*, pages 116–129, Porto Alegre, RS, Brasil. SBC.

Nockleby, J. (2000). Hate speech. In *Encyclopedia of the American Constitution*, page 1277–1279, New York, USA. Macmillan.

Pasqualotti, P. R. and Vieira, R. (2008). WordnetAffectBR: uma base lexical de palavras de emoções para a língua portuguesa. *RENOTE*, 6(1).

Silva, M. J., Carvalho, P., and Sarmento, L. (2012). Building a sentiment lexicon for social judgement mining. In Caseli, H., Villavicencio, A., Teixeira, A., and Perdigão, F., editors, *Computational Processing of the Portuguese Language*, pages 218–228, Berlin, Heidelberg. Springer Berlin Heidelberg.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Trajano, D., Bordini, R. H., and Vieira, R. (2023). OLID-BR: offensive language identification dataset for brazilian portuguese. *Lang. Resour. Eval.*, 58(4):1263–1289.

Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., and Benevenuto, F. (2022). HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.

Vargas, F., Rodrigues de Góes, F., Carvalho, I., Benevenuto, F., and Pardo, T. (2021). Contextual-lexicon approach for abusive language detection. In Mitkov, R. and Angelova, G., editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1438–1447, Held Online. INCOMA Ltd.