Multiple Voices, Greater Power: A Strategy for Combining Language Models to Combat Hate Speech

Annie Amorim‡, Gabriel Assis‡, Daniel de Oliveira, Aline Paes

Universidade Federal Fluminense, Niterói, Brasil
{annieamorim,assisgabriel}@id.uff.br, {danielcmo,alinepaes}@ic.uff.br

Abstract. Social media platforms face significant issues in avoiding a harmful environment with offensive comments and hate speech. Some of these challenges are inherently linked to the diversity of user perspectives, complicating the classification and detection of hate speech, particularly in culturally rich and diverse countries like Brazil. To address these complexities in identifying hate speech in Brazilian Portuguese, our work proposes the implementation of ensemble methods based on stacking and soft-voting, incorporating four distinct language models with varied architectures and pre-trainings: BERTimbau-base, BERTimbau-large, BERTweet.BR, and Bernice. The findings reveal the superiority of the proposed approach over individual prediction models, suggesting that the combination of multiple models may effectively integrate different perspectives, resulting in an accuracy improvement of up to 6% compared to the isolated classifications of the models.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**; **Ensemble methods**.

Keywords: ensemble, hate speech, social networks, transformers

1. INTRODUÇÃO

Expressões textuais são naturalmente diversas, uma vez que diferentes leitores podem perceber significados distintos a partir de um mesmo texto [Leonardelli et al. 2023]. Essa característica torna desafiante tarefas de classificação de texto, como a análise de sentimento [Kenyon-Dean et al. 2018] e a identificação de conteúdo ofensivo ou discurso de ódio [Assis et al. 2024]. Algumas abordagens consideram que tal subjetividade deve ser preservada, pois a diversidade de interpretações pode enriquecer a análise e proporcionar uma representação mais próxima da realidade [Leonardelli et al. 2023]. Especialmente no contexto de textos ofensivos e na detecção de discurso de ódio, variações linguísticas, expressões culturais e regionalismos podem gerar múltiplas interpretações na percepção desses conteúdos [Assis et al. 2024]. A compreensão dessas nuances se torna importante, portanto, para desenvolver abordagens mais eficazes e representativas na análise de textos dessa natureza.

Contudo, a presença de diversas perspectivas em conjuntos de dados representa um desafio significativo para a criação de modelos de aprendizado de máquina baseados em modelos de linguagem capazes de classificar instâncias textuais de maneira adequada. Para tratar desse desafio, emergem propostas como o *Learning with Disagreements* (Le-Wi-Di) [Uma et al. 2021], que se dedica ao aprendizado a partir de dados que contêm anotações subjetivas, incluindo, muitas vezes, aquelas contraditórias. Esse paradigma abrange, por exemplo, métodos que permitem o aprendizado diretamente a partir das

Os autores notificam que este artigo inclui exemplos de conteúdo ofensivo e de discurso de ódio, utilizados unicamente como ilustrações dos problemas discutidos, e que não expressam de forma alguma as suas opiniões ou visões pessoais. Além disso, agradecem ao financiamento do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), bolsa 307088/2023-5, da Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), processos SEI-260003/002930/2024, SEI-260003/000614/2023, e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Código Financeiro 001. ‡ Igual contribuição.

anotações fornecidas por diversos rotuladores ou que agregam as predições de múltiplos modelos para uma mesma instância [Leonardelli et al. 2023].

Ademais, as redes sociais enfrentam um crescente desafio em relação à propagação de conteúdo ofensivo e discurso de ódio [Aluru et al. 2020; Vargas et al. 2021]. Distinguir entre conteúdos ofensivos, discursos de ódio e textos neutros dos usuários nesse contexto é uma tarefa complexa, dada, por exemplo, a forte influência de figuras de linguagem, como a ironia, e o uso próprio da língua nos ambientes digitais. Especificamente, o Brasil, classificado como o terceiro maior consumidor de mídias sociais do mundo[1], representa um cenário particularmente desafiador devido a sua pluralidade. Tais fatores ressaltam a importância da avaliação do contexto e da subjetividade. Dessa forma, com o papel central das plataformas digitais na comunicação moderna, uma solução efetiva para este problema torna-se essencial [Saraiva et al. 2021].

Considerando esse cenário, este artigo avalia abordagens de agregação das probabilidades de saída geradas para cada classe por múltiplos modelos de linguagem, formando comitês (do inglês, *ensembles*) de classificadores [Zhou 2012]. Conjectura-se que a subjetividade e a discordância inerentes ao problema possam estar melhor capturadas nas saídas numéricas dos modelos, enquanto esses aspectos podem se perder na atribuição do rótulo final único. Para isso, utilizam-se dois conjuntos de dados: HateBR [Vargas et al. 2022] e ToLD-Br [Leite et al. 2020], e quatro modelos: BERTimbau-base, BERTimbau-large [Souza et al. 2020], BERTweet.BR [Caneiro et al. 2024] e Bernice [DeLucia et al. 2022], avaliando estratégias baseadas em *stacking* [Wolpert 1992], *soft-voting* e *hard-voting* [Zhou 2012]. Os resultados mostram que as abordagens de comitê que utilizam as probabilidades superam tanto as predições individuais dos modelos quanto a estratégia de *hard-voting*. Assim, este trabalho contribui com uma abordagem que integra múltiplas saídas no sensível domínio do discurso de ódio.

As seções deste artigo são divididas da seguinte forma: a Seção 2 discute trabalhos relacionados; a Seção 3 detalha a abordagem proposta; a Seção 4 explicita as configurações experimentais adotadas; a Seção 5 discute os resultados experimentais; e, por fim, a Seção 6 traz as considerações finais.

2. TRABALHOS RELACIONADOS

Diversos estudos sobre detecção e classificação de discurso de ódio destacam os classificadores baseados em modelos de linguagem derivados da arquitetura BERT [Devlin et al. 2019] como abordagens proeminentes para tais tarefas [Vargas et al. 2021; da Silva and Rosa 2023; Chu et al. 2024]. Assis et al. [2024] e Oliveira et al. [2024] concluem, inclusive, acerca de sua superioridade sobre os emergentes modelos de larga escala (LLMs) generativos. No entanto, essas abordagens não consideram explicitamente a subjetividade ou divergência entre anotadores durante o treinamento ou inferência, uma consideração recente e predominante no inglês [Leonardelli et al. 2023]. Diversos métodos utilizam agregação de múltiplos rótulos antes do treinamento ou na inferência. Akhtar et al. [2019] propõem manipulações nos dados que considerem perspectivas de múltiplos anotadores e medidas de polarização. Sullivan et al. [2023] sugerem representar cada anotador por uma camada linear adicional a um modelo de rede neural, agregando os resultados com uma camada extra ponderada.

Embora a aplicação de comitês seja frequentemente associada à algoritmos baseados em árvores [Zhou 2012], diferentes abordagens são adotadas nesse contexto. Pelle et al. [2018] propõem um comitê baseado em *word embeddings* para a detecção de textos ofensivos. Alguns estudos recentes combinam modelos de linguagem baseados no BERT, buscando maior generalização e robustez em domínios sensíveis. Shahriar and Solorio [2023], por exemplo, exploram dois comitês baseados em BERT: um que considera o aprendizado de cada anotador por um modelo diferente e outro que treina um modelo em uma tarefa de regressão para capturar a média agregada das anotações. Akhtar et al. [2021] propõem treinar classificadores distintos considerando subgrupos de anotadores por critérios

[1] `https://bit.ly/forbes-consumo-de-redes-br`

como raça, gênero e etnia, sugerindo um comitê em que a resposta final é positiva se pelo menos um classificador apontar para a classe positiva. Já Mnassri et al. [2022] avaliam a predição do modelo BERT combinada com outras arquiteturas, como CNNs e LSTMs, para a detecção de discurso de ódio. Ademais, Risch and Krestel [2020] propõem um *Bagging* [Breiman 1996] de múltiplas instâncias do mesmo modelo BERT para a tarefa de identificação de agressão em textos.

Nossa abordagem se distingue por empregar as probabilidades de saída geradas por diferentes modelos de linguagem — pré-treinados em conjuntos de dados distintos e com variedade arquitetural — ajustados para a classificação de discurso de ódio, partindo da conjectura de que esses modelos podem incorporar perspectivas distintas em suas saídas numéricas. A metodologia não exige anotações de múltiplos anotadores, porém reconhece a subjetividade inerente à detecção e classificação de conteúdo de ódio. Por fim, avalia-se também a aplicabilidade dessa estratégia no contexto do português brasileiro, investigando sua eficácia e adequação às particularidades linguísticas e culturais do Brasil.

3. COMITÊS DE MODELOS DE LINGUAGEM

Esta seção apresenta os modelos de linguagem selecionados e descreve as estratégias de agregação adotadas para integrar suas classificações, formando comitês. Foram aplicadas estratégias de agregação baseadas em *stacking* [Wolpert 1992], nas quais meta-modelos são treinados para combinar as predições dos modelos de linguagem, e uma abordagem de agregação de *soft-voting* [Zhou 2012], utilizando as probabilidades que cada modelo de linguagem atribui a cada classe diretamente. Ao combinar os resultados preditivos de diferentes modelos, busca-se incorporar os diversos conhecimentos codificados durante seus processos de pré-treinamento, proporcionando assim uma variedade de perspectivas para a classificação final de cada texto.

3.1 Modelos de Linguagem Selecionados

Ao encontro de Assis et al. [2024] e Oliveira et al. [2024], modelos de linguagem pré-treinados para o português ou em textos de redes sociais produzem os melhores resultados para detecção de discurso de ódio no idioma. Por esse motivo, quatro modelos foram selecionados: dois modelos pré-treinados com textos estruturados do português do Brasil, (i.) BERTimbau-base e (ii.) BERTimbau-large [Souza et al. 2020]; (iii.) BERTweet.BR [Caneiro et al. 2024], pré-treinado com *tweets* brasileiros; e (iv.) Bernice [DeLucia et al. 2022], um modelo multilíngue, também pré-treinado com *tweets*.

3.2 Estratégias de Agregação

Para combinar o aprendizado de múltiplos modelos de linguagem ajustados para classificação textual, utilizando o conhecimento incorporado em seus pesos, foi empregada a técnica de comitê denominada *stacking*. Essa abordagem organiza os modelos em dois níveis: modelos-base e meta-modelo. Inicialmente, os modelos-base são treinados de forma independente, e suas previsões são utilizadas como entrada para o treinamento dos modelos do segundo nível. O meta-modelo, então, é treinado para realizar a previsão final com base no espaço de *features* gerado pelos modelos anteriores. Em termos gerais, a função do meta-modelo é aprender a combinar eficientemente as previsões dos modelos-base. Nesse sentido, os modelos de linguagem foram aplicados como modelos-base, e meta-modelos distintos foram avaliados, especificamente usando as abordagens *Random Forest*, *Logistic Regression* e *Support Vector Machine* (SVM).

Durante a fase de treinamento do *stacking*, cria-se um novo conjunto de dados a partir dos modelos-base para treinar os meta-modelos. Isso ajuda a mitigar o risco de sobreajuste que ocorreria no caso de, em ambos os níveis do *stacking*, o mesmo conjunto de treinamento fosse utilizado. Optou-se por utilizar um processo fundamentado na validação cruzada estratificada *k-fold* [Zhou 2012]. Nesse processo, o **conjunto de treinamento original** D é dividido em k partes iguais $(D_1, \ldots, D_k)$.

Definem-se então D_i e $D_{(-i)} = D \setminus D_i$, que são os conjuntos de teste e treinamento para o i-ésimo *fold*, respectivamente. Em cada passo i, m **modelos de linguagem** são treinados sobre o conjunto $D_{(-i)}$ e geram inferências sobre a partição D_i, que é isolada em cada iteração. Ao final, as predições sobre as k partições isoladas D_i são combinadas para formar o novo conjunto de dados necessário. Como resultado, as variações no treinamento de cada *fold* implicam o espaço de *features* gerado não ser uma captura única do conjunto de treinamento original, podendo ser assim usado para o treinamento dos meta-modelos. A Figura 1 ilustra esse processo. Por fim, com o novo conjunto de treinamento para o segundo nível de modelos gerados, os modelos de linguagem base são retreinados sobre todo o conjunto de treinamento original D. Reforça-se que, nesse contexto, a inferência realizada por cada modelo de linguagem resulta na geração de *soft-labels*. Esses *soft-labels* compõem uma c-upla $\{p_1, p_2, \ldots, p_c\}$, onde c é o número máximo de classes possíveis para o problema de classificação.

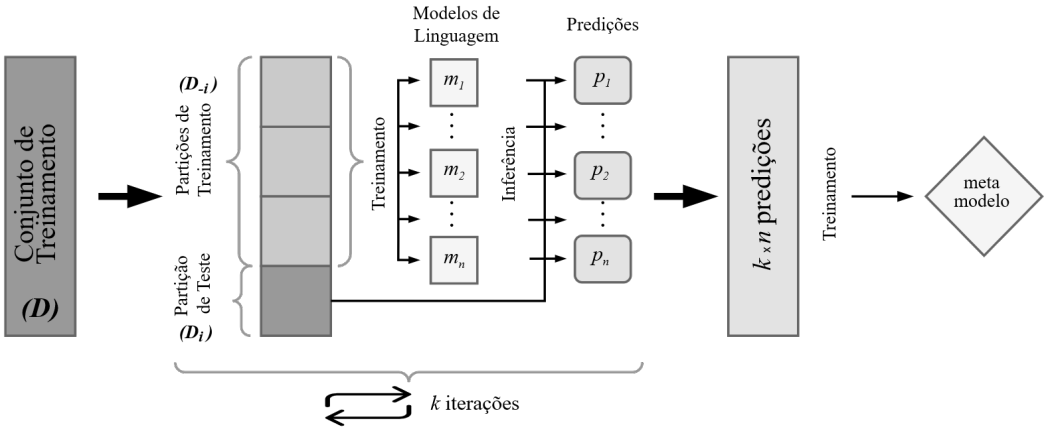

Fig. 1. Treinamento do meta-modelo para o *stacking*.

Adicionalmente, avaliou-se também uma abordagem mais simples de *soft-voting* para combinar as classificações dos modelos de linguagem. Nela, a previsão final é determinada pelo máximo entre as médias das **probabilidades para cada classe** p_c por todos os modelos de linguagem. Em termos gerais, enquanto na abordagem de *stacking* utiliza-se uma função de aprendizado que tenta capturar a melhor combinação de pesos para cada modelo, nessa abordagem mais simples, todos os modelos possuem o mesmo peso. Essa estratégia de agregação será tratada no texto como *Simple Average Probability Voting* (SAPV). Em síntese, a classe predita para cada **instância de texto** x, segundo essa abordagem, é definida simplesmente pela equação:

$$c_x = \arg\max_c \left(\frac{1}{m} \sum_{i=1}^{m} p_c^i(x) \right) \tag{1}$$

Ao fim, para avaliações e predições, as inferências dos comitês são geradas ao se passar as instâncias de texto pelos modelos de linguagem, resultando nas probabilidades atribuídas a cada classe. Essas probabilidades são então agregadas utilizando-se os meta-modelos treinados ou pela estratégia de SAPV, conforme descrito anteriormente. Ademais, avaliou-se comparativamente o *hard-voting*, em que a classe final é determinada pela maioria entre as classes finais atribuídas pelos modelos-base. Em casos de empate, adotou-se a classe de maior severidade como vencedora [Akhtar et al. 2021].

4. CONFIGURAÇÃO EXPERIMENTAL

Esta seção apresenta os conjuntos de dados utilizados e os detalhes de implementação.

4.1 Conjuntos de Dados

Dois conjuntos de dados com conteúdo de ódio foram utilizados para a avaliação: HateBR [Vargas et al. 2022], que inclui comentários em perfis de políticos brasileiros no Instagram, e ToLD-Br [Leite et al. 2020], que consiste em uma coleção de *tweets* brasileiros. Os conjuntos foram obtidos pré-processados conforme disponível publicamente por Assis et al. [2024]. Para o HateBR, 5.417 instâncias constituem treino, enquanto 1.358 foram reservadas para teste. Para o ToLD-Br, esses números são 16.150 e 4.015, respectivamente. O pré-processamento dos textos envolveu a anonimização de usuários com o *token* @USER, a substituição de URLs pelo *token* HTTPURL e a conversão de *emojis* em suas representações textuais. Adicionalmente, também consoante Assis et al. [2024], devido ao desequilíbrio de classes, optou-se por utilizar a versão balanceada do conjunto de treinamento. Por fim, os conjuntos estavam processados para um problema de classificação ternário entre as classes {neutro, ofensivo, discurso de ódio}, consideradas, nessa ordem, em termos de severidade.

4.2 Configurações de Hiperparâmetros e Detalhes de Implementação

Conforme Assis et al. [2024], os modelos de linguagem foram ajustados por meio de *fine-tuning* completo, utilizando o *framework transformers*[2], com $epochs = 2$ e $learning\_rate = 2 \cdot 10^{-5}$. O mesmo *framework* foi empregado para as inferências. Já os meta-modelos foram instanciados com a biblioteca scikit-learn[3] nas seguintes configurações: *RandomForestClassifier(n_estimators=100, criterion='log_loss')*, *SVC(kernel='rbf')* e *LogisticRegression(max_iter=100, multi_class='ovr', penalty='l2')*. O processo apresentado na Seção 3.2 foi aplicado com cinco *folds (k=5)*. Visando garantir a reprodutibilidade, todos os parâmetros de *random_state* foram configurados para o valor 42.

5. RESULTADOS EXPERIMENTAIS

Esta seção apresenta os resultados experimentais, abordando a eficácia dos modelos de classificação no domínio do discurso de ódio por meio de métricas de classificação e de uma inspeção qualitativa.

5.1 Resultados Classificatórios

Tabela I. Resultados classificatórios das abordagens avaliadas. Os valores destacados em negrito representam os melhores resultados obtidos, enquanto os valores sublinhados correspondem aos segundos melhores desempenhos.

Modelo/Estratégia	**HateBR**				**ToLD-Br**			
	acc.	**prec.**	**rec.**	**F1**	**acc.**	**prec.**	**rec.**	**F1**
Modelos Individuais								
BERTimbau-base	0,7953	0,7228	0,7805	0,7358	0,6406	0,4839	0,5364	0,4743
BERTimbau-large	0,8358	0,7655	0,8279	0,7834	0,6443	0,5241	0,6286	0,4968
Bernice	0,7791	0,7134	0,7714	0,7152	0,6105	0,4800	0,5365	0,4577
BERTweet.BR	0,8351	0,7628	0,8226	0,7799	0,6481	0,5126	0,6546	0,5009
Comitês								
Logistic Regression	**0,8490**	**0,7782**	**0,8454**	**0,7961**	0,6742	0,5305	0,6598	0,5171
Random Forest	0,8292	0,7590	0,8277	0,7742	0,6695	0,5229	0,6552	0,5111
SVM	0,8439	0,7729	0,8383	0,7891	0,6777	0,5312	0,6694	0,5213
SAPV	**0,8490**	0,7744	0,8345	0,7914	**0,6862**	**0,5340**	**0,6695**	**0,5284**
Hard-voting	0,8159	0,7473	0,8173	0,7589	0,6585	0,5237	0,6385	0,5054

A Tabela I compara o desempenho dos comitês com o desempenho individual dos modelos que os compõem. Observa-se que, entre os modelos individuais, o BERTimbau-large obteve as melhores métricas no conjunto HateBR e apresentou o melhor desempenho de precisão no conjunto ToLD-Br. Por outro lado, o BERTweet.BR, um modelo pré-treinado em *tweets* em português, destacou-se nas outras três métricas avaliadas nesse último conjunto. Por sua vez, estratégias de agregação das

[2]https://huggingface.co/docs/transformers/
[3]https://scikit-learn.org/

predições probabilísticas dos quatro modelos superaram os desempenhos individuais e o *hard-voting* em ambos os conjuntos de dados, destacando-se *Logistic Regression* para o HateBR e o SAPV para o ToLD-Br. Como ilustração, observou-se um aumento de 6% na acurácia e de 5% no valor de F1 no conjunto ToLD-Br ao comparar a melhor predição individual com o melhor comitê. Esses resultados evidenciam a superioridade das abordagens de comitê propostas e destacam o potencial de considerar as probabilidades de saída em comparação ao *hard-voting*. Dessa forma, contribui-se para um desempenho mais robusto e eficaz, o que é importante considerando o contexto subjetivo do discurso de ódio e a sensibilidade dos erros nesse cenário. Falsos positivos podem resultar em censura indevida, enquanto falsos negativos podem falhar em proteger grupos identitários.

Contudo, nenhuma estratégia de comitê se destaca como a melhor de forma absoluta. Por exemplo, a *Logistic Regression* se mostrou mais eficaz para o conjunto de dados HateBR, enquanto a SAPV obteve os melhores resultados para o ToLD-Br. Ademais, a SVM utilizada como meta-modelo de agregação apresentou desempenhos positivos em ambos os conjuntos. Já a estratégia SAPV também alcançou resultados comparáveis aos melhores no conjunto HateBR. Casos interessantes se dão sobre a *Random Forest* e o *hard-voting*, também nesse conjunto, em que suas aplicações geram resultados inferiores até mesmo ao uso individual do BERTimbau-large e BERTweet.BR. Assim, embora as estratégias de comitê sejam promissoras, é importante avaliá-las em cada contexto específico.

5.2 Inspeção Qualitativa

A Tabela II apresenta exemplos das classificações das estratégias avaliadas. Primeiro, destaca-se que a abordagem de comitê utilizando *Logistic Regression* (C1) apresentou o melhor desempenho em termos de F1 para o conjunto de dados HateBR. Dentre as 139 instâncias rotuladas como discurso de ódio, essa estratégia falhou na classificação em apenas 22 delas. Esses erros, em sua maioria, consistiram na classificação equivocada de discursos de ódio como textos ofensivos, destacando a complexidade na diferenciação entre essas duas classes. Essa questão é exemplificada pelo texto 2 da tabela, que, inclusive, evidencia uma divisão nas respostas dos modelos. Além disso, alguns dos textos contêm expressões ambíguas, como ilustrado pelo texto 1, em que a *hashtag* "#pepapig" pode ser interpretada tanto com uma conotação gordofóbica quanto como referência direta ao desenho infantil. Adicionalmente, nota-se a capacidade do comitê C1 em classificar corretamente casos em que os modelos individuais não obtiveram sucesso, exemplificados nos textos 3 e 4. No entanto, na classe de textos ofensivos, o C1 apresentou algumas discrepâncias. Em particular, houve 20 casos em que todos os modelos, incluindo comitês, classificaram os textos como ofensivos, enquanto o C1 os classificou como discurso de ódio. Esses casos se dividem entre aqueles em que, aparentemente, não há discurso de ódio de fato, como no texto 6, e situações em que a definição é mais desafiadora, como no texto 7, que pode ser interpretado como contendo teor xenofóbico.

No conjunto ToLD-Br, o comitê SAPV (C4) teve o melhor resultado. Dos 58 casos rotulados como discurso de ódio, o comitê classificou incorretamente 22. Alguns desses erros são instigantes, não em função de falhas nos modelos, mas devido à rotulação original dos dados. Por exemplo, no texto 8, o termo "bichinha" pode ter sido associado a discurso de ódio, porém, em algumas regiões do Brasil, ele pode ser usado de maneira carinhosa, o que aparenta ser o caso. Já o texto 9 contém um termo vulgar, mas que não constitui discurso de ódio. Por outro lado, alguns casos de acerto, como nos textos 10 e 11, revelam-se particularmente interessantes. Apesar da divergência nas respostas dos modelos individuais, o comitê classificou corretamente. Isso reflete sua capacidade de agregar as probabilidades de saída de cada modelo em suas predições, demonstrando o potencial da combinação de múltiplos modelos representando diferentes perspectivas.

Destaca-se ainda que o *hard-voting* (C5) falha em cenários como os dos textos 4 e 11, em que ao menos um comitê que considera as probabilidades classifica corretamente. Isso pode ser um reflexo da sua menor capacidade de captar nuances, que podem estar presentes nas probabilidades, as quais são ignoradas na agregação desse método.

Tabela II. **R**ótulos e predições dos modelos BERTimbau-base (**M1**), BERTimbau-large (**M2**), Bernice (**M3**) e BERTweet.BR (**M4**), e dos comitês baseados em *Logistic Regression* (**C1**), *Random Forest* (**C2**), SVM (**C3**), SAPV (**C4**) e *Hard-voting* (**C5**) nos **D**ados do **H**ateBR e do **T**oLD-Br. Denota-se neutro(**0**), ofensivo(**1**) e discurso de ódio(**2**).

D		Texto	R	M1	M2	M3	M4	C1	C2	C3	C4	C5
H	1	#pepapig	2	2	2	0	1	1	2	2	2	2
H	2	Esse bolsolixo é um canalha...	2	2	1	1	2	1	1	1	1	2
H	3	Uma das mais privilegiadas mentes do meio jornalístico. Não trouxe o rancor da globolixo. Parabéns Leda.	1	0	0	0	2	1	0	0	0	0
H	4	Enquanto um protege nossas riquezas, o outro se preocupa com o cú alheio	2	1	1	1	1	2	1	1	1	1
H	5	Que escroto caquético! É a velha hipocrisia judaica no mundo dos pilantras monetários.	2	2	2	2	2	1	2	2	2	2
H	6	Sem novidades, lixo sendo lixo.	1	1	1	1	1	2	1	1	1	1
H	7	É uma absurdo né? Não sei o que fazem aqui... deviam estar em Cuba ou Venezuela...	1	1	1	1	1	2	1	1	1	1
T	8	eu tô só adiando pra ir na casa de vovó, a bichinha já nem acredita quando falo que vou lá	2	1	0	0	0	0	0	0	0	0
T	9	tô cheia de fome e nada desse ônibus vir, que merda!	2	1	0	1	1	0	1	1	1	1
T	10	e ainda tem essa de "bi de balada" aí sinceramente vcs são nojentos	2	2	1	2	1	2	2	2	2	2
T	11	@USER kkkk minha frase eu coloquei sua bixa	2	1	0	2	0	2	2	2	2	0

6. CONSIDERAÇÕES FINAIS

Este estudo investiga a implementação de comitês baseados nas probabilidades de saída de múltiplos modelos de linguagem para a classificação de textos no domínio do discurso de ódio. Os resultados indicam que a abordagem proposta supera a execução individual dos modelos e o *hard-voting*. Ao incorporar múltiplas visões no processo de aprendizado e inferência, representadas nas saídas de diversos modelos, busca-se contribuir para a literatura recente, que enfatiza a necessidade de considerar e tratar a subjetividade inerente ao discurso de ódio. Para pesquisas futuras, pretende-se expandir a avaliação para novos conjuntos de dados subjetivos, além de comparar a aplicabilidade do método com modelos de linguagem de larga escala e em cenários com rótulos fornecidos por múltiplos anotadores.

REFERÊNCIAS

Akhtar, S., Basile, V., and Patti, V. A new measure of polarization in the annotation of hate speech. In *AI*IA 2019 – Advances in Artificial Intelligence*, M. Alviano, G. Greco, and F. Scarcello (Eds.). Springer International Publishing, Cham, pp. 588–603, 2019.

Akhtar, S., Basile, V., and Patti, V. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection, 2021.

Aluru, S. S., Mathew, B., Saha, P., and Mukherjee, A. Deep learning models for multilingual hate speech detection, 2020.

Assis, G., Amorim, A., Carvalho, J., de Oliveira, D., Vianna, D., and Paes, A. Exploring Portuguese hate speech detection in low-resource settings: Lightly tuning encoder models or in-context learning of large models? In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, and R. Amaro (Eds.). Association for Computational Lingustics, Santiago de Compostela, Galicia/Spain, pp. 301–311, 2024.

Breiman, L. Bagging predictors. *Machine Learning* 24 (2): 123–140, Aug, 1996.

Caneiro, F., Viana, D., Carvalho, J., Plastino, A., and Paes, A. BERTweet.BR: A Pre-Trained Language Model for Tweets in Portuguese. *Neural Computing and Applications*, 2024. Accepted, to appear.

Chu, T. M., Weitzel, L., and Quaresma, P. Comparative analysis of hate speech detection models on brazilian portuguese data: Modified bert vs. bert vs. standard machine learning algorithms. In *Proceedings of the 13th International Conference on Data Science, Technology and Applications - Volume 1: DATA*. INSTICC, SciTePress, Dijon, France, pp. 392–400, 2024.

da Silva, R. C. C. and Rosa, T. C. Combining data transformation and classification approaches for hate speech detection: A comparative study. Available at SSRN, 2023.

DeLucia, A., Wu, S., Mueller, A., Aguirre, C., Resnik, P., and Dredze, M. Bernice: A Multilingual Pre-trained Encoder for Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 6191–6205, 2022.

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

for Computational Linguistics: Human Language Technologies, Volume 1, J. Burstein, C. Doran, and T. Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186, 2019.

Kenyon-Dean, K., Ahmed, E., Fujimoto, S., Georges-Filteau, J., Glasz, C., Kaur, B., Lalande, A., Bhanderi, S., Belfer, R., Kanagasabai, N., Sarrazingendron, R., Verma, R., and Ruths, D. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, M. Walker, H. Ji, and A. Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, pp. 1886–1895, 2018.

Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. 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*, K.-F. Wong, K. Knight, and H. Wu (Eds.). Association for Computational Linguistics, Suzhou, China, pp. 914–924, 2020.

Leonardelli, E., Abercrombie, G., Almanea, D., Basile, V., Fornaciari, T., Plank, B., Rieser, V., Uma, A., and Poesio, M. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori (Eds.). Association for Computational Linguistics, Toronto, Canada, pp. 2304–2318, 2023.

Mnassri, K., Rajapaksha, P., Farahbakhsh, R., and Crespi, N. BERT-based ensemble approaches for hate speech detection. In *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*. Institute of Electrical and Electronics Engineers, Rio de Janeiro, Brazil, pp. 4649–4654, 2022.

Oliveira, A., de Carvalho Cecote, T., Alvarenga, J. P. R., de Souza Freitas, V. L., and da Silva Luz, E. J. Toxic Speech Detection in Portuguese: A Comparative Study of Large Language Models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, and R. Amaro (Eds.). Association for Computational Lingustics, Santiago de Compostela, Galicia/Spain, pp. 108–116, 2024.

Pelle, R., Alcântara, C., and Moreira, V. P. A classifier ensemble for offensive text detection. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*. WebMedia '18. Association for Computing Machinery, New York, NY, USA, pp. 237–243, 2018.

Risch, J. and Krestel, R. Bagging BERT models for robust aggression identification. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, R. Kumar, A. K. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, and D. Kadar (Eds.). European Language Resources Association (ELRA), Marseille, France, pp. 55–61, 2020.

Saraiva, G. D., Anchiêta, R., Neto, F. A. R., and Moura, R. A semi-supervised approach to detect toxic comments. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. INCOMA Ltd., Held Online, pp. 1261–1267, 2021.

Shahriar, S. and Solorio, T. SafeWebUH at SemEval-2023 task 11: Learning annotator disagreement in derogatory text: Comparison of direct training vs aggregation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori (Eds.). Association for Computational Linguistics, Toronto, Canada, pp. 94–100, 2023.

Souza, F., Nogueira, R., and Lotufo, R. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems*, R. Cerri and R. C. Prati (Eds.). Springer International Publishing, Cham, pp. 403–417, 2020.

Sullivan, M., Yasin, M., and Jacobs, C. L. University at buffalo at SemEval-2023 task 11: MASDA–modelling annotator sensibilities through DisAggregation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori (Eds.). Association for Computational Linguistics, Toronto, Canada, pp. 978–985, 2023.

Uma, A., Fornaciari, T., Dumitrache, A., Miller, T., Chamberlain, J., Plank, B., Simpson, E., and Poesio, M. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, and X. Zhu (Eds.). Association for Computational Linguistics, Online, pp. 338–347, 2021.

Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., and Benevenuto, F. HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis (Eds.). European Language Resources Association, Marseille, France, pp. 7174–7183, 2022.

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

Wolpert, D. H. Stacked generalization. *Neural Networks* 5 (2): 241–259, 1992.

Zhou, Z.-H. *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC, 2012.