Leveraging Task-Adaptive Pretraining for Misogynistic Hate Speech Classification in Brazilian Portuguese

Ana Rubia R. Vicente[1], **Hilário T. A. de Oliveira**[1],
Kelly A. de S. Gazolli[1], **Karin S. Komati**[1]

[1]Programa de Pós-graduação em Computação Aplicada (PPComp)
Instituto Federal do Espírito Santo (IFES)
Serra – ES – Brasil

anarubiarvicente@gmail.com, {hilario.oliveira,kasouza,kkomati}@ifes.edu.br

Abstract. *This work analyzes the impact of Task-Adaptive Pretraining (TAPT) on the identification of misogynistic hate speech in Brazilian Portuguese. To this end, the MINA-BR dataset was used, with the unlabeled portion used for TAPT and the annotated portion for fine-tuning and evaluation. The study compares the performance of the RoBERTa, ALBERT, BERTimbau, and BERTabaporu architectures via cross-validation. The results indicate that TAPT improves performance in all architectures, particularly the fMINABr-BERTabaporu model. These findings suggest that TAPT improves language models in scenarios with scarce annotated data and high linguistic diversity.*

Resumo. *Este trabalho analisa o impacto do pré-treinamento adaptado à tarefa (TAPT) na identificação de discurso de ódio misógino em português do Brasil. Para isso, foi utilizado o conjunto de dados MINA-BR, cuja porção não rotulada é empregada no TAPT, enquanto a anotada é utilizada para ajuste fino e avaliação. O trabalho compara o desempenho das arquiteturas RoBERTa, ALBERT, BERTimbau e BERTabaporu via validação cruzada. Os resultados indicam que o TAPT melhora o desempenho em todas as arquiteturas, com destaque para o modelo fMINABr-BERTabaporu. Esses achados sugerem que o TAPT contribui para o aprimoramento de modelos de linguagem em cenários com escassez de dados anotados e alta diversidade linguística.*

1. Introdução

Discurso de ódio pode ser definido como qualquer manifestação oral, textual ou por meio de comportamentos que tenha como alvo indivíduos ou grupos, utilizando linguagem ofensiva ou discriminatória baseada em aspectos ligados à identidade [de Pelle and Moreira 2017]. Quando direcionado a mulheres ou meninas, recebe a denominação de discurso de ódio misógino, caracterizando-se por mensagens que perpetuam desigualdades de gênero, expressando hostilidade, preconceito ou incentivando atitudes de menosprezo [Saha et al. 2018, Ramos and da Silva Figueiredo 2018].

No contexto brasileiro, a identificação automática desse tipo de discurso apresenta desafios adicionais, entre os quais se destacam a escassez de bases de dados anotadas em português e a ampla diversidade linguística observada nas interações, especialmente em ambientes *online*, como as redes sociais. Embora iniciativas como a MINA-BR [Plath et al. 2022, Plath et al. 2025] representem avanços relevantes, a quantidade ainda

limitada de dados rotulados por humanos impõe restrições ao treinamento de modelos de aprendizado profundo, que tipicamente exigem grandes volumes de exemplos para alcançar boa capacidade de generalização.

Diante da diferença de escala frequentemente observada entre grandes volumes de dados textuais brutos e a quantidade limitada de dados anotados disponíveis para o treinamento supervisionado, uma alternativa adotada na literatura para melhor aproveitar dados não rotulados é o Pré-treinamento Adaptado à Tarefa (TAPT, do inglês *Task-Adaptive Pretraining*). Essa técnica consiste em realizar uma etapa adicional de pré-treinamento em modelos de linguagem previamente treinados em corpora de domínio geral, utilizando um conjunto de textos não rotulados relacionado à tarefa de interesse [Gururangan et al. 2020]. No processo de TAPT, o modelo é exposto a dados do domínio específico desejado, o que possibilita ajustar suas representações internas às características linguísticas, semânticas e terminológicas presentes nesse contexto. Como resultado, o modelo passa a capturar padrões e vocabulários específicos diretamente do corpus não rotulado antes da etapa de ajuste fino supervisionado.

Este trabalho tem como objetivo avaliar a eficácia da aplicação do método de TAPT na identificação automática de discurso de ódio em textos escritos em português do Brasil. Para isso, é realizada uma análise comparativa do desempenho de quatro modelos baseados na arquitetura *Bidirectional Encoder Representations from Transformers* (BERT) [Devlin et al. 2019], incluindo modelos com pré-treinamento em domínio geral (RoBERTa, ALBERT, BERTimbau) e um modelo especializado em textos do Twitter (BERTabaporu). Os experimentos consideram dois cenários distintos, com e sem a aplicação de uma etapa adicional de pré-treinamento utilizando o método TAPT.

O processo de ajuste fino é realizado seguindo as diretrizes descritas em [Grotti and Quick 2024] e utiliza a metodologia de validação cruzada *k-fold*, o que permite analisar a estabilidade dos hiperparâmetros e o desempenho dos modelos em diferentes partições de dados de treinamento e validação. A avaliação final é conduzida em um conjunto de dados independente, com o objetivo de mensurar a capacidade de generalização dos modelos em exemplos inéditos que não foram utilizados nas etapas de treinamento ou validação.

2. Trabalhos Relacionados

A detecção automática de discurso de ódio evoluiu de abordagens baseadas em regras e atributos linguísticos manuais para métodos que utilizam representações distribuídas de palavras, como Word2Vec e GloVe. Mais recentemente, modelos baseados na arquitetura *Transformers*, como BERT e RoBERTa, tornaram-se predominantes devido ao desempenho superior obtido com pré-treinamento em grandes corpora e melhor adaptação a diferentes domínios e idiomas [Rawat et al. 2024].

O trabalho de [Grotti and Quick 2024], desenvolvido para a tarefa HaSpeeDe 3 do EVALITA 2023, investiga a aplicação de modelos de linguagem baseados no BERT, especificamente o UmBERTo e o *Italian BERT* (BERT-ita), na detecção de discurso de ódio para a língua italiana. A metodologia aplicada inicia-se com uma etapa de ajuste fino (*fine-tuning*) utilizando validação cruzada *k-fold* ($k = 5$), seguida de uma avaliação final em conjuntos de dados independentes (*holdout*). Os resultados demonstram que o modelo UmBERTo obteve desempenho superior à média dos participantes da competição.

O uso de técnicas de adaptação tem sido explorado em línguas com poucos recursos, que apresentam características semelhantes ao cenário brasileiro. Belay et al. [Belay et al. 2025] demonstram que a aplicação de TAPT e *Domain-Adaptive Continual Pretraining* (DAPT) pode reduzir o viés de modelos pré-treinados em línguas africanas, os quais frequentemente apresentam dificuldades para representar adequadamente a linguagem dinâmica presente nas mídias sociais devido à predominância de dados provenientes de domínios genéricos ou religiosos durante o pré-treinamento original.

Trabalhos voltados ao português brasileiro têm demonstrado que modelos de linguagem pré-treinados alcançam desempenho consistente em tarefas de detecção de discurso de ódio. O modelo BERTimbau, por exemplo, foi utilizado em bases como OFFCOMBR-2 e ORCOMBR-3, obtendo resultados superiores quando combinado com estratégias de pré-processamento e balanceamento por *oversampling* [Silva and Freitas 2022]. De maneira semelhante, Ramos de Oliveira et al. [Ramos de Oliveira et al. 2023] investigaram diferentes variantes do modelo BERTimbau, nas versões *base* e *large*, observando desempenhos comparáveis entre elas, o que sugere que arquiteturas mais compactas podem obter resultados competitivos para esse tipo de tarefa. Em [Assis et al. 2024], experimentos conduzidos com modelos do tipo *encoder*, incluindo BERTimbau e Albertina PT-BR, indicam que mesmo estratégias de ajuste fino com poucas épocas de treinamento podem produzir resultados competitivos na detecção de discurso de ódio.

O trabalho de [Plath et al. 2022] apresenta a construção da base MINA-BR, voltada à identificação de discurso de ódio contra mulheres em português brasileiro. Os autores aplicaram técnicas de pré-processamento textual e utilizaram representações baseadas em *Bag-of-Words* (BoW) e *Term Frequency-Inverse Document Frequency* (TF-IDF), avaliando classificadores tradicionais, como *Support Vector Machine* (SVM), Floresta Aleatória e variantes do *Naive Bayes*. O melhor resultado foi obtido com SVM utilizando TF-IDF sem *stemming*, alcançando um valor de F1-score de 0,57.

Vicente et al. [Vicente et al. 2025] investigaram diferentes técnicas de aprendizado supervisionado para a detecção automática de discurso de ódio misógino utilizando o conjunto de dados MINA-BR. Os autores compararam abordagens baseadas em representações textuais tradicionais, como BoW e TF-IDF, com o modelo pré-treinado BERTimbau, empregando classificadores como *Naive Bayes*, SVM, Regressão Logística e *Multi-Layer Perceptron* (MLP). Os resultados indicam que o BERTimbau apresentou o melhor desempenho geral, alcançando 87% de acurácia e F1-score de 0,58 para a classe de ódio, superando os métodos tradicionais.

Além dos avanços em desempenho, estudos recentes também têm enfatizado aspectos relacionados à interpretabilidade e à construção de recursos linguísticos especializados. Nesse contexto, Salles et al. [Salles et al. 2025] propõem a base HateBRXplain, um conjunto de dados de referência para a detecção explicável de discurso de ódio em português brasileiro. O estudo avaliou modelos, como o BERTimbau e o PTT5, utilizando métodos agnósticos como LIME e SHAP para testar a interpretabilidade das decisões automatizadas. Os resultados indicam que, embora os modelos alcancem alto desempenho na classificação, eles ainda enfrentam desafios para gerar explicações fiéis e plausíveis.

Embora os trabalhos demonstrem a eficácia de modelos pré-treinados para a

detecção de discurso de ódio em diferentes idiomas e domínios, poucos estudos investigam o impacto da adaptação adicional do modelo ao domínio específico da tarefa por meio do uso de dados não rotulados. Diferentemente dos trabalhos supracitados, este trabalho avalia sistematicamente a aplicação do método de TAPT, utilizando os exemplos não anotados da coleção MINA-BR, explorando uma característica ainda pouco investigada na literatura de discurso de ódio em português brasileiro. Além disso, são comparadas diferentes arquiteturas baseadas em BERT, incluindo modelos treinados em domínio geral e modelos especializados em redes sociais, utilizando um protocolo experimental baseado na metodologia de validação cruzada *k-fold* e avaliação final em um conjunto de teste independente.

3. Materiais e Métodos

3.1. Base de dados

A base de dados MINA-BR [Plath et al. 2025, Plath et al. 2022], utilizada neste trabalho, contém 6.002 comentários extraídos de redes sociais, incluindo YouTube e Twitter, coletados a partir de palavras-chave relacionadas ao tema. Do total de comentários disponíveis, 2.091 foram anotados manualmente por especialistas quanto à presença ou ausência de discurso de ódio misógino, enquanto os 3.911 registros restantes não possuem rótulos. Neste trabalho, foram utilizados dois campos da base para cada instância: (i) `comment`, correspondente ao texto completo do comentário; e (ii) `odio`, que indica a classificação binária atribuída, indicando Ódio misógino (1) ou Não ódio misógino (0).

A distribuição das classes no corpus MINA-BR Rotulada é composta por 340 exemplos da classe Ódio misógino e 1.751 exemplos de comentários que não contêm discurso de ódio. Essa distribuição evidencia um desbalanceamento significativo entre as categorias, condição recorrente em tarefas de detecção de discurso de ódio [De Gibert et al. 2018]. Para os experimentos, o corpus MINA-BR Rotulada foi dividido de forma estratificada em conjuntos de treino e teste, na proporção de 90% (1.882 exemplos) e 10% (209 exemplos), respectivamente.

3.2. Abordagens Investigadas

Os experimentos propostos neste trabalho contemplam duas abordagens distintas. A primeira abordagem, ilustrada na Figura 1, consiste no ajuste fino (*fine-tuning*) direto dos modelos de linguagem pré-treinados. O processo inicia-se com os modelos RoBERTa, ALBERT, BERTimbau e BERTabaporu. Para cada modelo, é adicionada uma camada de classificação. Em seguida, os modelos são ajustados para a tarefa-alvo utilizando a base MINA-BR rotulada. Para aumentar a robustez da avaliação, foi utilizada a metodologia de validação cruzada k-fold (k = 5). Após o treinamento, os modelos resultantes, identificados pelo prefixo "f-", são avaliados em um conjunto de teste separado do mesmo corpus, permitindo a obtenção das medidas finais de desempenho. O código-fonte dos experimentos encontra-se disponível em um repositório do GitHub[1].

A segunda abordagem, ilustrada na Figura 2, incorpora uma etapa intermediária de *Task-Adaptive Pretraining* (TAPT) antes do processo de *fine-tuning*. Inicialmente, os

[1] `https://github.com/anarvicente/eniac`

modelos base (RoBERTa, ALBERT, BERTimbau e BERTabaporu) passam por um pré-treinamento adicional utilizando o conjunto composto por 3.911 comentários não rotulados extraídos da base MINA-BR. Esse processo segue o paradigma de *continual pre-training*, no qual os parâmetros do modelo são atualizados por meio da função objetivo de *Masked Language Modeling* (MLM). Durante essa etapa, *tokens* das sentenças são parcialmente ocultados e o modelo é treinado para prever os termos mascarados a partir do contexto textual, permitindo adaptar o vocabulário e as representações semânticas ao domínio da tarefa. Os modelos resultantes desse processo são identificados pelo prefixo "fMINABr-". Após essa etapa de adaptação, o fluxo experimental segue o mesmo procedimento adotado na primeira abordagem.

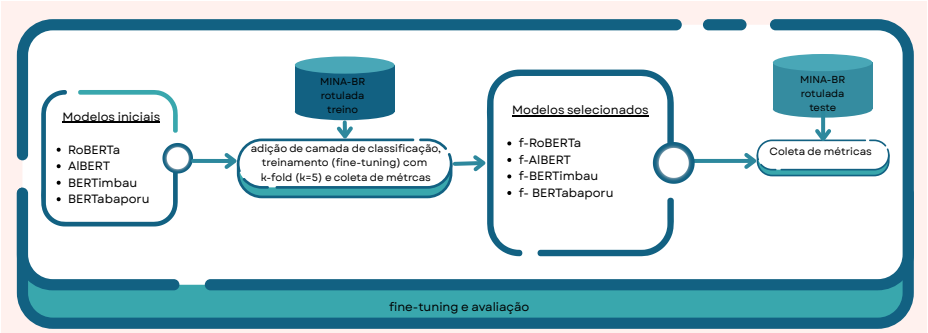


Figura 1. Fluxograma de detecção de discurso de ódio com *fine-tuning*.

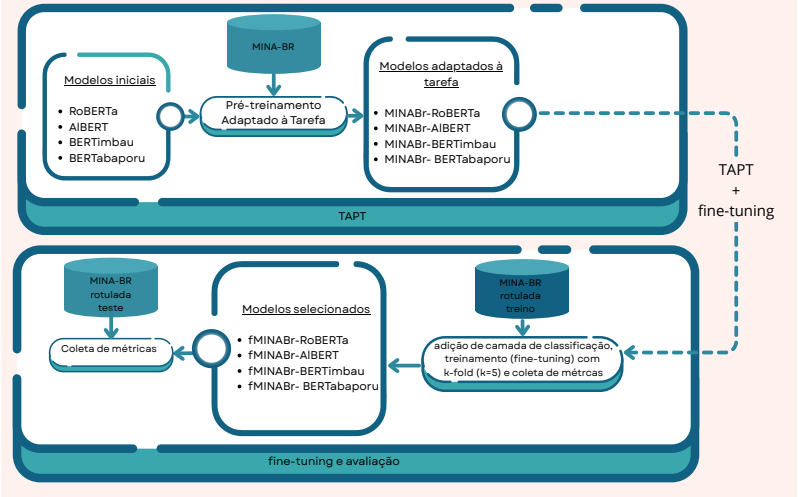


Figura 2. Fluxograma de detecção de discurso de ódio com TAPT e *fine-tuning*.

Os experimentos deste trabalho utilizam modelos baseados na arquitetura Transformer da família *Bidirectional Encoder Representations from Transformers* (BERT) [Devlin et al. 2019]. O BERT é um modelo de linguagem pré-treinado que emprega codificadores *Transformer* bidirecionais para capturar dependências contextuais profundas no texto. Com base nessa arquitetura, foram avaliados quatro modelos amplamente utilizados em tarefas de processamento de linguagem natural: RoBERTa [Liu et al. 2019], ALBERT [Lan et al. 2019], BERTimbau Base [Souza et al. 2020] e BERTabaporu [Costa et al. 2023]. O RoBERTa é uma variação do BERT que otimiza o processo de pré-treinamento por meio de ajustes como a remoção da tarefa de predição

de próxima sentença e o uso de mascaramento dinâmico. O ALBERT introduz estratégias de compartilhamento de parâmetros e fatoração de *embeddings*, reduzindo significativamente o número de parâmetros do modelo. Já o BERTimbau consiste em uma adaptação do BERT pré-treinada especificamente para o português brasileiro, utilizando o corpus brWaC. Por sua vez, o BERTabaporu foi treinado com um grande volume de *tweets* em português, permitindo capturar características linguísticas típicas de redes sociais, como informalidade, abreviações e variações ortográficas.

3.3. Pré-processamento e Refinamento dos Modelos

Para a abordagem que incorpora a etapa intermediária de TAPT, cada arquitetura passou inicialmente pelo processo de tokenização correspondente ao seu modelo base. Em seguida, os textos foram segmentados em sequências com comprimento máximo de 128 *tokens*. Durante o pré-treinamento adicional, usando como função objetivo a MLM, uma fração dos *tokens* foi mascarada aleatoriamente, permitindo que os modelos aprendessem a prever palavras ocultas com base no contexto textual.

O processo de pré-treinamento foi conduzido utilizando o otimizador AdamW, com taxa de aprendizado inicial de 5×10^{-5} e esquema de decaimento linear. O treinamento foi realizado por três épocas, utilizando o *framework Accelerate*, que permite melhor gerenciamento de recursos computacionais em ambientes de treinamento distribuído. Para avaliar o desempenho dos modelos durante essa etapa, foi utilizada a métrica de perplexidade, amplamente empregada em tarefas de modelagem de linguagem para medir a qualidade da distribuição probabilística aprendida pelo modelo. Após o treinamento, os modelos BERTimbau e BERTabaporu apresentaram perplexidade de 14,00, seguidos pelo RoBERTa com 15,00. O ALBERT apresentou desempenho inferior nesse cenário, registrando perplexidade de 62,00, indicando menor capacidade de adaptação ao corpus utilizado. Ao final dessa etapa, os modelos refinados foram armazenados em repositórios próprios na plataforma Hugging Face[2].

Na etapa de *fine-tuning*, comum nas duas abordagens adotadas, a base de dados foi particionada de forma estratificada em 90% para treinamento e 10% para teste. O subconjunto de teste foi mantido como conjunto *hold-out* para a avaliação final, enquanto o conjunto de treinamento foi utilizado no processo de validação cruzada. O treinamento foi realizado com tamanho de lote (*batch size*) de 16 exemplos e semente aleatória fixada em 42, garantindo a reprodutibilidade dos experimentos. Em cada *fold*, os modelos foram treinados por 15 épocas, sendo realizadas etapas de avaliação ao final de cada época. O modelo com melhor desempenho (F1 Macro) no conjunto de validação foi selecionado para representar cada execução experimental. Em todos os experimentos foram utilizados os mesmos hiperparâmetros de treinamento para todos os modelos de linguagem, assegurando a comparabilidade direta entre as arquiteturas avaliadas.

4. Resultados e Discussão

4.1. Resultados do *fine-tuning* sem TAPT

Na Tabela 1 são apresentados os resultados obtidos por meio de validação cruzada de 5-*fold* para a tarefa de classificação, sem a aplicação de TAPT. Na tabela são apresentados os

[2] https://huggingface.co/

valores médios e os desvios padrão das cinco iterações da validação cruzada. Os melhores resultados obtidos em cada medida de avaliação estão destacados em negrito.

Tabela 1. Resultados da validação cruzada (5-*folds*) sem TAPT.

Modelo	Classe/Média	Precisão	Revocação	F1-score	Suporte
RoBERTa	Classe "Não Ódio"	$0,92 \pm 0,015$	$0,92 \pm 0,018$	$0,92 \pm 0,007$	315
	Classe "Ódio"	$0,58 \pm 0,042$	$0,58 \pm 0,092$	$0,58 \pm 0,052$	61
	Média Macro	$0,75 \pm 0,023$	$0,75 \pm 0,041$	$0,75 \pm 0,028$	376
	Média Ponderada	$0,86 \pm 0,015$	$0,86 \pm 0,012$	$0,86 \pm 0,013$	376
ALBERT	Classe "Não Ódio"	$0,91 \pm 0,003$	$0,92 \pm 0,026$	$0,91 \pm 0,012$	315
	Classe "Ódio"	$0,57 \pm 0,092$	$0,51 \pm 0,018$	$0,54 \pm 0,031$	61
	Média Macro	$0,74 \pm 0,046$	$0,72 \pm 0,011$	$0,72 \pm 0,023$	376
	Média Ponderada	$0,85 \pm 0,017$	$0,85 \pm 0,020$	$0,85 \pm 0,016$	376
BERTimbau	Classe "Não Ódio"	$0,91 \pm 0,009$	$0,93 \pm 0,016$	$0,92 \pm 0,007$	315
	Classe "Ódio"	$0,61 \pm 0,046$	$0,55 \pm 0,058$	$0,57 \pm 0,038$	61
	Média Macro	$0,76 \pm 0,025$	$0,74 \pm 0,026$	$0,75 \pm 0,022$	376
	Média Ponderada	$0,86 \pm 0,011$	$0,87 \pm 0,012$	$0,87 \pm 0,012$	376
BERTabaporu	Classe "Não Ódio"	$\mathbf{0,93 \pm 0,013}$	$\mathbf{0,93 \pm 0,031}$	$\mathbf{0,93 \pm 0,011}$	315
	Classe "Ódio"	$\mathbf{0,65 \pm 0,084}$	$\mathbf{0,64 \pm 0,082}$	$\mathbf{0,63 \pm 0,026}$	61
	Média Macro	$\mathbf{0,79 \pm 0,037}$	$\mathbf{0,78 \pm 0,044}$	$\mathbf{0,78 \pm 0,016}$	376
	Média Ponderada	$\mathbf{0,88 \pm 0,011}$	$\mathbf{0,88 \pm 0,016}$	$\mathbf{0,88 \pm 0,011}$	376

O modelo BERTabaporu atingiu os índices macro de precisão $(0,79)$ e F1-score $(0,78)$, valores superiores aos demais modelos avaliados. O resultado sugere que o pré-treinamento em dados de redes sociais constitui uma vantagem relevante para a tarefa investigada, já que o corpus MINA-BR é composto por comentários extraídos de plataformas como YouTube e Twitter. Essa proximidade entre o domínio de pré-treinamento e o domínio de aplicação tende a favorecer a captura de características linguísticas informais, como abreviações, variações ortográficas e construções mais coloquiais, frequentes em interações online. Em contraste, o ALBERT apresentou o desempenho inferior entre as arquiteturas, com F1-score macro de $0,72$, o que pode estar relacionado ao seu tamanho de arquitetura mais compacto e à redução de parâmetros por compartilhamento entre camadas. Embora essa estratégia seja eficiente do ponto de vista computacional, ela pode limitar a capacidade do modelo de representar nuances semânticas em uma tarefa sensível ao contexto, como a detecção de discurso de ódio.

As arquiteturas RoBERTa e BERTimbau demonstraram índices próximos na métrica de F1-score macro, ambas com valor de $0,75$. No entanto, o BERTimbau apresentou ligeira vantagem no equilíbrio entre precisão e revocação para a classe 1 ("Ódio"), o que sugere que o pré-treinamento em português brasileiro contribui positivamente para a tarefa, ainda que não seja suficiente, por si só, para superar um modelo treinado especificamente em linguagem de redes sociais.

A classe 0 ("Não Ódio") apresenta índices de F1-score entre $0,91$ e $0,93$ em todas as arquiteturas testadas. No entanto, o desempenho na classe 1 ("Ódio") revela a complexidade da tarefa de detecção sob condições de desbalanceamento de dados. O BER-

Tabaporu obteve um valor de F1-score de $0,63$ para a classe minoritária, índice acima dos valores registrados pelo RoBERTa $(0,58)$, BERTimbau $(0,57)$ e ALBERT $(0,54)$. Os desvios padrão observados nas cinco iterações indicam estabilidade razoável entre as execuções dos modelos. Isso sugere que os resultados não decorrem de uma partição particularmente favorável dos dados, mas refletem um comportamento relativamente consistente das arquiteturas ao longo da validação cruzada.

4.2. Resultados do *fine-tuning* com TAPT

Na Tabela 2 são apresentados os resultados (média $\pm$ desvio padrão) obtidos via validação cruzada de 5-*fold* usando o método TAPT para os modelos fMINABr-RoBERTa, fMINABr-ALBERT, fMINABr-BERTimbau e fMINABr-BERTabaporu. A aplicação da técnica de pré-treinamento adaptado à tarefa resultou em alterações nos índices de desempenho das arquiteturas. O modelo BERTabaporu registrou um F1-score macro de $0,80$, mantendo a posição de destaque entre os modelos avaliados. O ALBERT apresentou F1-score macro de $0,74$, enquanto as arquiteturas BERTimbau e RoBERTa alcançaram índices de $0,77$ e $0,76$, respectivamente. A inclusão do TAPT elevou o desempenho na identificação da classe 1 em todas as arquiteturas em comparação ao cenário de ajuste fino direto. O BERTabaporu atingiu F1-score macro de $0,66$, um acréscimo em relação ao valor de $0,63$ obtido sem a adaptação. O BERTimbau também apresentou ganho nesta métrica, atingindo o valor de $0,62$ (ante $0,57$), enquanto o RoBERTa e o ALBERT registraram $0,60$ e $0,56$. Os índices para a classe 0 permaneceram em patamares entre $0,92$ e $0,93$ em todas as configurações.

Tabela 2. Resultados da validação cruzada (5-*folds*) com o método de TAPT.

Modelo	Classe/Média	Precisão	Revocação	F1-score	Suporte
fMINABr-RoBERTa	Classe "Não Ódio"	$0,92 \pm 0,010$	$0,93 \pm 0,026$	$0,92 \pm 0,015$	315
	Classe "Ódio"	$0,62 \pm 0,097$	$0,58 \pm 0,056$	$0,60 \pm 0,061$	61
	Média Macro	$0,77 \pm 0,048$	$0,75 \pm 0,031$	$0,76 \pm 0,037$	376
	Média Ponderada	$0,87 \pm 0,021$	$0,87 \pm 0,026$	$0,87 \pm 0,023$	376
fMINABr-ALBERT	Classe "Não Ódio"	$0,91 \pm 0,014$	$0,93 \pm 0,014$	$0,92 \pm 0,006$	315
	Classe "Ódio"	$0,60 \pm 0,032$	$0,53 \pm 0,078$	$0,56 \pm 0,055$	61
	Média Macro	$0,75 \pm 0,018$	$0,73 \pm 0,037$	$0,74 \pm 0,026$	376
	Média Ponderada	$0,86 \pm 0,015$	$0,86 \pm 0,010$	$0,86 \pm 0,013$	376
fMINABr-BERTimbau	Classe "Não Ódio"	$0,92 \pm 0,006$	$\mathbf{0,94 \pm 0,016}$	$\mathbf{0,93 \pm 0,006}$	315
	Classe "Ódio"	$\mathbf{0,66 \pm 0,049}$	$0,58 \pm 0,040$	$0,62 \pm 0,019$	61
	Média Macro	$0,79 \pm 0,025$	$0,76 \pm 0,015$	$0,77 \pm 0,011$	376
	Média Ponderada	$0,88 \pm 0,007$	$0,88 \pm 0,009$	$0,88 \pm 0,007$	376
fMINABr-BERTabaporu	Classe "Não Ódio"	$\mathbf{0,94 \pm 0,011}$	$0,93 \pm 0,009$	$\mathbf{0,93 \pm 0,005}$	315
	Classe "Ódio"	$\mathbf{0,66 \pm 0,026}$	$\mathbf{0,67 \pm 0,066}$	$\mathbf{0,66 \pm 0,040}$	61
	Média Macro	$\mathbf{0,80 \pm 0,016}$	$\mathbf{0,80 \pm 0,030}$	$\mathbf{0,80 \pm 0,022}$	376
	Média Ponderada	$\mathbf{0,89 \pm 0,011}$	$\mathbf{0,89 \pm 0,009}$	$\mathbf{0,89 \pm 0,010}$	376

A comparação entre os cenários indica que a inclusão de uma etapa a mais de pré-treinamento em conjunto de exemplos não rotulados no mesmo domínio da tarefa favorece a captura de padrões da classe 1. No BERTabaporu, o equilíbrio entre precisão $(0,66)$ e

revocação $(0, 67)$ reflete a estabilidade da predição após o uso de TAPT. Os desvios padrão indicam a manutenção da consistência dos resultados durante a validação cruzada, com valores de dispersão em faixas reduzidas nas métricas de F1. Esses resultados corroboram a eficácia da adaptação à tarefa para a detecção de nuances em contextos de escassez de anotação.

4.3. Resultados na base de testes

A Tabela 3 apresenta os resultados dos modelos com e sem a aplicação de TAPT no conjunto de teste. Os modelos identificados pelo prefixo "f-" correspondem ao fluxo de *fine-tuning* sem TAPT, enquanto aqueles com o prefixo "fMINABr-" incluem uma etapa prévia de TAPT. Os melhores resultados em cada medida de avaliação estão destacados em negrito. Os resultados no conjunto de teste independente (*holdout*) corroboram as tendências observadas na validação cruzada, indicando boa capacidade de generalização dos modelos. De modo geral, a aplicação de TAPT produziu ganhos consistentes de desempenho, especialmente na classe minoritária. Esses resultados sugerem que a adaptação ao corpus não rotulado permitiu especializar as representações do modelo às características linguísticas e semânticas do domínio antes da etapa de *fine-tuning* supervisionado.

Tabela 3. Desempenho dos modelos na base de teste MINA-BR Rotulada.

Modelo	Precisão (0)	Revocação (0)	F1 (0)	Precisão (1)	Revocação (1)	F1 (1)
f-RoBERTa	0,9096	0,9148	0,9122	0,5455	0,5294	0,5373
f-ALBERT	0,8962	**0,9318**	**0,9136**	**0,5556**	0,4412	0,4918
f-BERTimbau	0,9006	0,9261	0,9132	0,5517	0,4706	0,5079
f-BERTabaporu	**0,9345**	0,8920	0,9128	0,5476	**0,6765**	**0,6053**
fMINABr-RoBERTa	0,9106	0,9261	0,9183	0,5806	0,5294	0,5538
fMINABr-ALBERT	0,8756	**0,9602**	0,9160	0,5882	0,2941	0,3922
fMINABr-BERTimbau	0,9034	0,9034	0,9034	0,5000	0,5000	0,5000
fMINABr-BERTabaporu	**0,9419**	0,9205	**0,9310**	**0,6316**	**0,7059**	**0,6667**

Considerando os resultados sem o uso do TAPT, o f-BERTabaporu obteve o melhor F1 na classe 1 $(0, 6053)$, especialmente devido à revocação de $0, 6765$, superior aos modelos f-RoBERTa $(0, 5294)$, f-BERTimbau $(0, 4706)$ e f-ALBERT $(0, 4412)$. A inclusão da etapa de pré-treinamento extra de TAPT elevou o desempenho, destacando-se o fMINABr-BERTabaporu, que alcançou um F1-score de $0, 6667$ na classe 1. Este ganho foi acompanhado por maior precisão $(0, 6316)$ e revocação $(0, 7059)$, indicando que a adaptação ao MINA-BR aumentou a detecção de verdadeiros positivos, controlando os falsos positivos. O fMINABr-RoBERTa também evoluiu, atingindo um F1-score de $0, 5538$.

A comparação evidencia que o TAPT especializa as arquiteturas para o domínio de redes sociais. Enquanto o fMINABr-ALBERT regrediu na classe 1 (F1-score de $0, 3922$), o BERTabaporu consolidou sua generalização. A estabilidade na classe 0, com F1-score entre $0, 90$ e $0, 93$, confirma que a especialização não prejudicou a classificação de textos neutros. Esse resultado reforça a hipótese de que modelos previamente alinhados ao domínio de redes sociais tendem a se beneficiar mais intensamente da adaptação adicional à tarefa.

Analisando os resultados do modelo fMINABr-BERTabaporu em comparação com a literatura, é possível observar uma evolução nos índices de classificação de discurso de ódio na base MINA-BR. O trabalho de [Plath et al. 2022] registra um F1-score de $0,57$ com o uso de classificadores SVM e representação usando TF-IDF. O estudo de [Vicente et al. 2025] reporta um F1-score de $0,58$ para a classe de ódio com o emprego da arquitetura BERTimbau. Em contrapartida, o fMINABr-BERTabaporu atinge um F1-score de $0,6667$ na classe 1. Os valores demonstram que a aplicação de pré-treinamento adaptado à tarefa supera os resultados de abordagens exploradas em trabalhos anteriores usando algoritmos de aprendizado de máquina tradicionais e de ajuste fino de modelos de linguagem em português.

Outro aspecto é que o desempenho da classe majoritária permaneceu alto e relativamente estável em todas as arquiteturas, com F1-score entre $0,92$ e $0,93$. Isso sugere que a introdução do TAPT não comprometeu a capacidade de classificação das instâncias de "Não Ódio", mas produziu ganhos principalmente na classe de maior interesse prático. O caso do ALBERT merece atenção particular. Apesar de apresentar desempenho competitivo na classe majoritária, o modelo sofreu deterioração na classe 1 após o TAPT, atingindo F1-score de $0,3922$ no conjunto *holdout*. Esse comportamento sugere que a arquitetura não conseguiu se beneficiar da adaptação ao domínio e possivelmente apresenta menor capacidade de capturar as nuances linguísticas necessárias para a identificação do discurso misógino no contexto analisado. A combinação entre pior perplexidade na etapa de MLM e pior desempenho final na classe minoritária reforça essa interpretação.

De forma geral, os resultados com o conjunto de teste separado *holdout* sustentam a principal conclusão do estudo de que a inclusão de uma etapa extra de pré-treinamento com TAPT apresenta melhores resultados quando aplicada a modelos cujo pré-treinamento apresenta afinidade com o domínio discursivo da tarefa. O desempenho do BERTabaporu sugere que a adequação do corpus de pré-treinamento original continua sendo um fator central, e que o TAPT atua como uma etapa complementar de especialização, e não como substituto para um pré-treinamento bem alinhado ao domínio.

5. Conclusão

Este trabalho apresentou uma avaliação de modelos de linguagem baseados na arquitetura BERT para a detecção de discurso de ódio misógino em português brasileiro, utilizando o conjunto de dados MINA-BR. Os resultados indicaram que o pré-treinamento adaptado à tarefa (TAPT) esteve associado a melhorias no desempenho dos modelos avaliados. Entre as arquiteturas analisadas, o fMINABr-BERTabaporu apresentou os melhores resultados, superando valores reportados em trabalhos anteriores sobre a mesma base. Esses resultados indicam o potencial do uso de dados não rotulados para a adaptação de modelos de linguagem em cenários de baixa disponibilidade de dados rotulados.

Apesar dos resultados promissores, eles devem, contudo, ser interpretados considerando as limitações da base MINA-BR, que possui apenas 340 exemplos rotulados da classe de ódio e um conjunto de teste *holdout* com poucas instâncias positivas. Essa característica pode aumentar a sensibilidade das métricas e a variabilidade dos resultados. Além disso, a etapa extra de pré-treinamento de TAPT foi realizada por poucas épocas e com poucos dados de treinamento, o que pode ter limitado a sua eficácia. Como trabalhos futuros, pretende-se avaliar diferentes volumes de dados não rotulados no TAPT, inves-

tigar técnicas de aumento de dados e comparar modelos baseados em codificadores com modelos de linguagem de larga escala, investigar a generalização dos modelos em diferentes domínios e plataformas digitais, considerando variações linguísticas e contextuais.

Agradecimentos

Os autores agradecem ao CNPq pelo projeto 407742/2022-0 e à FAPES pelos projetos 1023/2022 (P:2022-8TZV6), 055/2026 (P: 2026-6FNG4) e 374/2026 (P: 2026-XMJSB). Profª. Komati agradece ao CNPq pela bolsa DT-2 (302726/2023-3).

Referências

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

Belay, T. D., Azime, I. A., Ahmad, I. S., Adelani, D. I., Abdulmumin, I., Ayele, A. A., Muhammad, S. H., and Yimam, S. M. (2025). AfroXLMR-social: Adapting pretrained language models for African languages social media text. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V., editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15570–15587, Suzhou, China. Association for Computational Linguistics.

Costa, P. B., Pavan, M. C., Santos, W. R., Silva, S. C., and Paraboni, I. (2023). BERTabaporu: Assessing a genre-specific language model for Portuguese NLP. In Mitkov, R. and Angelova, G., editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 217–223, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

De Gibert, O., Perez, N., García-Pablos, A., and Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

de Pelle, R. and Moreira, V. (2017). Offensive comments in the Brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*, pages 510–519, Porto Alegre, RS, Brasil. SBC.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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 (long and short papers)*, pages 4171–4186.

Grotti, L. and Quick, P. (2024). Berticelli at haspeede 3: Fine-tuning and cross-validating large language models for hate speech detection. In *EVALITA Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian Final Workshop: Parma, Italy, September 7-8th, 2023*, volume 2(3), page 188. Accademia University Press.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and

tasks. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv:cs.CL*, page 1907.11692.

Plath, H. O., Paiva, M. E. O., Pinto, D. L., and Costa, P. D. P. (2022). Detecção de discurso de Ódio contra mulheres em textos em português brasileiro: Construção da base mina-br e modelo de classificação. *Revista Eletrônica de Iniciação Científica em Computação*, 20(3).

Plath, H. O., Paiva, M. E. O., Pinto, D. L., and Costa, P. D. P. (2025). MINA-BR: base de dados para identificação de discurso de ódio contra a mulher na internet. https://doi.org/10.25824/redu/MQD68Z.

Ramos, S. B. and da Silva Figueiredo, K. (2018). Uma taxonomia dos tipos de preconceito enfrentados por mulheres na Área de tecnologia. In *Anais do XII Women in Information Technology*, pages 91–94, Porto Alegre, RS, Brasil. SBC.

Ramos de Oliveira, F., Reis, V., and Ebecken, N. (2023). Tupy-e: detecting hate speech in brazilian portuguese social media with a novel dataset and comprehensive analysis of models. *arXiv preprint arXiv:2312.17704*.

Rawat, A., Kumar, S., and Samant, S. S. (2024). Hate speech detection in social media: Techniques, recent trends, and future challenges. *Wiley Interdisciplinary Reviews: Computational Statistics*, 16(2):e1648.

Saha, P., Mathew, B., Goyal, P., and Mukherjee, A. (2018). Hateminers: Detecting hate speech against women. *arXiv preprint arXiv:1812.06700*.

Salles, I., Vargas, F., and Benevenuto, F. (2025). Hatebrxplain: A benchmark dataset with human-annotated rationales for explainable hate speech detection in brazilian portuguese. In *Proceedings of the 31st international conference on computational linguistics*, pages 6659–6669.

Silva, F. and Freitas, L. (2022). Brazilian portuguese hate speech classification using bertimbau. *The International FLAIRS Conference Proceedings*, 35.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417, Cham. Springer International Publishing.

Vicente, A. R. R., Andrade, J. O., de Souza Gazolli, K. A., and Komati, K. S. (2025). Comparative analysis of machine learning models for misogynistic hate speech detection in brazilian portuguese: A study using the mina-br dataset. In *Proceedings of the IEEE Chilecon 2025 Conference, 2025, Valparaíso, Chile. 2025*.