Detecção de Viés Ideológico em Artigos de Notícias Utilizando Aprendizagem Métrica Profunda e Representações Contextuais

Jailson Pereira Januário, André Luiz da Costa Carvalho

[1]Instituto de Computação
Universidade Federal do Amazonas (UFAM) – Manaus, AM – Brasil

{jailson,acarvalho}@icomp.ufam.edu.br

Abstract. *This study investigates ideological bias detection in news articles using deep metric learning with DistilRoBERTa. In the best experimental configuration, the model achieved a Macro F1-Score of 0.83 on ABP under random split. Under the more rigorous media-bias split, performance dropped to 0.43, highlighting the challenge of cross-source generalization. In zero-shot comparison, the proposed approach surpassed GPT-3.5 and Llama-2 in Macro F1, indicating better class balance for this task. Overall, the results support metric learning as a competitive strategy for ideological bias detection while exposing important limitations in domain adaptation across news ecosystems.*

Resumo. *Este estudo investiga a detecção de viés ideológico em artigos de notícias por meio de aprendizagem métrica profunda com DistilRoBERTa. Na melhor configuração experimental, o modelo atingiu Macro F1-Score de 0,83 no conjunto ABP em random split. Sob o cenário mais rigoroso de media-bias split, o desempenho caiu para 0,43, evidenciando os desafios de generalização entre fontes. Na comparação com LLMs em regime zero-shot, a abordagem proposta superou GPT-3.5 e Llama-2 em Macro F1, indicando melhor equilíbrio entre classes para esta tarefa. Em síntese, os resultados sustentam a aprendizagem métrica como estratégia competitiva para detecção de viés ideológico, ao mesmo tempo em que expõem limitações relevantes de adaptação de domínio entre ecossistemas de notícias.*

1. Introdução

Nos últimos anos, a expansão de informações online por portais de notícias intensificou o desafio de avaliar a imparcialidade dos conteúdos. O viés ideológico em artigos jornalísticos pode distorcer a percepção pública e influenciar decisões políticas e resultados eleitorais [Gentzkow and Shapiro 2006, Chiang and Knight 2011]. Dada a subjetividade inerente ao discurso político, a identificação automática dessa inclinação permanece uma tarefa complexa e de alta relevância para a integridade informacional.

No contexto de Processamento de Linguagem Natural (PLN), diferentes abordagens têm sido propostas para detectar viés com base em conteúdo textual, hiperlinks e teoria da informação [Spinde et al. 2021, Patricia Aires et al. 2019]. Entretanto, parte das soluções disponíveis ainda depende de fontes externas de metadados ou é avaliada em cenários excessivamente polarizados, o que limita sua aplicabilidade em domínios reais e heterogêneos. Além disso, modelos de linguagem de larga escala (LLMs), embora

competitivos em diversas tarefas e capazes de bons resultados em *zero-shot*, tendem a demandar maior custo computacional e operacional em uso contínuo. Em paralelo, modelos codificadores leves, como o DistilRoBERTa, podem alcançar desempenho comparável quando treinados em bases alinhadas ao domínio-alvo [Lin et al. 2024, Lin et al. 2025].

Para mitigar essas limitações, este trabalho investiga uma abordagem baseada em *Deep Metric Learning* para detecção de viés ideológico, com ênfase no equilíbrio entre desempenho e custo computacional. Em vez de depender de modelos generativos de maior porte, a proposta otimiza a geometria do espaço de representação com DistilRoBERTa, aproximando exemplos semanticamente similares e afastando exemplos dissimilares, de modo a viabilizar classificação em larga escala.

As contribuições deste estudo são:

- Propor um pipeline de detecção de viés ideológico baseado em *embeddings* contextuais e aprendizagem métrica profunda;
- Avaliar o impacto de *Contrastive Loss* e *Triplet Loss* com *semi-hard negative mining* na qualidade das representações;
- Analisar a generalização cruzada entre domínios por meio de experimentos entre diferentes conjuntos de dados;
- Comparar o desempenho da abordagem proposta com baselines da literatura e modelos LLM em regime *zero-shot*.

Para apresentar o que se propõe, este artigo está organizado como segue. Os trabalhos relacionados são discutidos na Seção 2. Os materiais e métodos são apresentados na Seção 3. Os resultados e a discussão encontram-se dispostos na Seção 4. Por fim, as considerações finais, limitações e futuros são apresentados na Seção 5.

2. Trabalhos Relacionados

A literatura sobre detecção de viés ideológico abrange diferentes perspectivas, desde abordagens baseadas em redes estruturais [Efron 2004] até modelos econômicos de reputação midiática [Gentzkow and Shapiro 2006]. Trabalhos posteriores integraram análise de redes sociais e PLN para quantificar inclinação política por meio de grafos de interação [Lin et al. 2011]. Contudo, tais estratégias tendem a perder robustez em cenários com poucos hiperlinks, ausência de relações explícitas entre fontes ou baixa conectividade entre documentos.

Outra linha de pesquisa utiliza metadados de plataformas sociais para inferir posicionamento ideológico de veículos e conteúdos. Apesar de útil em contextos com alta disponibilidade de dados externos, essa dependência de APIs e sinais comportamentais reduz a autonomia do método e dificulta sua aplicação em ambientes onde apenas o texto bruto está disponível.

No âmbito da análise textual direta, estudos exploram escolhas lexicais, recorrência temática e *framing* em manchetes como indicadores de viés [Dallmann et al. 2015, Gangula et al. 2019]. Modelos baseados em *Transformers*, como BERT e variantes, frequentemente atingem bom desempenho em domínios conhecidos, mas ainda apresentam degradação ao serem aplicados a fontes não vistas [Baly et al. 2020].

Com foco nessa lacuna, o framework *POLITICS* [Liu et al. 2022] introduziu o pré-treinamento de modelos RoBERTa com *triplet loss*, utilizando exemplos semanticamente relacionados para organizar o espaço de representação. Mais recentemente, abordagens especializadas com ajuste fino em múltiplos *datasets* demonstraram ganhos consistentes de robustez em classificação de viés [Volf and Simko 2025].

Ainda assim, tanto modelos especializados quanto LLMs podem apresentar vieses políticos inerentes ou desalinhamentos em cenários *zero-shot* [Lin et al. 2024, Lin et al. 2025]. Diante desse contexto, este trabalho adota *Deep Metric Learning* com *semi-hard sampling* para induzir representações discriminativas e operacionalmente eficientes, mantendo como hipótese que a cobertura do domínio de treinamento é fator central para desempenho em novas fontes. A principal diferença em relação a abordagens puramente classificatórias está em otimizar explicitamente a topologia do espaço vetorial, favorecendo a separação entre classes ideológicas mesmo em fontes inéditas.

3. Material e Métodos

A metodologia assume que o viés ideológico se manifesta em padrões discursivos, processando textos via modelos pré-treinados e otimizando *embeddings* por *Contrastive* e *Triplet Loss*. Conforme a Figura 1, o fluxo abrange quatro etapas: definição dos conjuntos de dados, geração de representações por aprendizagem métrica, treinamento de classificadores e análise de desempenho. As seções seguintes detalham cada módulo da arquitetura proposta.

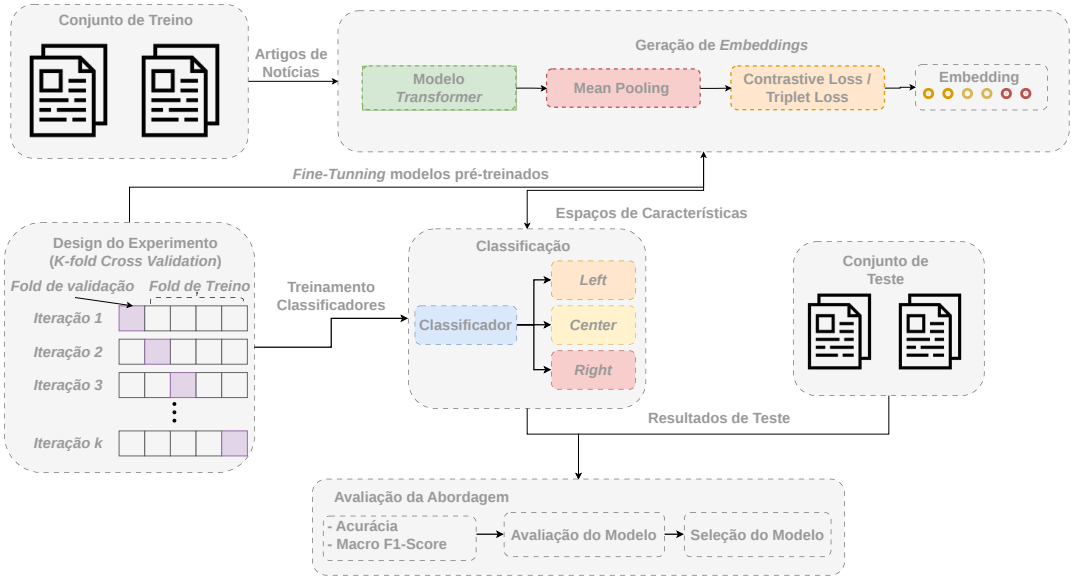


Figura 1. Visão geral do método de detecção de viés ideológico por meio do conteúdo textual de artigos de notícias.

3.1. Dados Experimentais

Para o desenvolvimento deste estudo, utilizaram-se os conjuntos de dados detalhados na Tabela 1. O foco principal recai sobre o *Article Bias Prediction* (ABP) [Baly et al. 2020], devido ao seu maior volume de amostras, à disponibilidade de rótulos ideológicos em três classes e à presença de partições previamente disponibilizadas pelo próprio conjunto de

dados, as quais controlam a distribuição de fontes jornalísticas entre treino, validação e teste.

Tabela 1. Resumo dos conjuntos de dados e volume de amostras.

Conjunto de Dados	Classes	Quantidade
FlipBias [Chen et al. 2018]	Left/Center/Right	3.066
ABP [Baly et al. 2020]	Left/Center/Right	36.274

A robustez metodológica foi validada por dois critérios de particionamento: o *random split* (Tabela 3), que utiliza amostragem aleatória para manter a consistência na distribuição de classes, e o *media-bias split* (Tabela 2). Este último segrega os artigos de forma que os dados de uma mesma fonte sejam exclusivos a uma única partição, mitigando o vazamento de dados e avaliando a generalização para veículos não vistos. Para garantir esse isolamento entre treino e teste, segue-se o protocolo de [Baly et al. 2020], descartando os artigos que excederiam o balanceamento por veículo no conjunto de teste; por essa razão, o total de amostras no *media-bias split* (30.246) é inferior ao total original do conjunto de dados (36.274).

Tabela 2. Estatísticas da partição *media-bias split* do conjunto de dados ABP.

Treino			**Validação**			**Teste**		
Viés	*Total*	*%*	*Viés*	*Total*	*%*	*Viés*	*Total*	*%*
Left	8.861	33,32%	Left	1.640	69,60%	Left	402	30,92%
Center	7.488	28,16%	Center	618	26,23%	Center	299	23,00%
Right	10.241	38,51%	Right	98	4,15%	Right	599	46,07%

Tabela 3. Estatísticas da partição *random split* do conjunto de dados ABP.

Treino			**Validação**			**Teste**		
Viés	*Total*	*%*	*Viés*	*Total*	*%*	*Viés*	*Total*	*%*
Left	9.750	34,84%	Left	2.438	34,84%	Left	402	30,92%
Center	7.988	28,55%	Center	1.998	28,55%	Center	299	23,00%
Right	10.240	36,60%	Right	2.560	36,59%	Right	599	46,07%

Adicionalmente, o experimento foi estendido para um cenário de generalização externa, onde o modelo treinado integralmente no conjunto ABP foi submetido à avaliação no conjunto *FlipBias* [Chen et al. 2018]. Este conjunto contém 2.781 eventos sob múltiplas perspectivas e serve como referência para medir a eficácia das representações em contextos fora do domínio de treinamento, seguindo as mesmas diretrizes estabelecidas por Lin et al. [2025].

3.2. Tarefa de Geração de *Embeddings*

Para a execução da tarefa, foi utilizado um modelo fundamentado no *Bidirectional Encoder Representations from Transformers (BERT)* [Devlin et al. 2018], reconhecidos pela eficácia na modelagem de dependências de longo alcance e na extração de relações semânticas granulares [Zhang and Rao 2020]. A seleção recaiu sobre o DistilRoBERTa

[Liu et al. 2019], variante destilada que preserva a robustez da arquitetura original, contudo, apresenta reduções substanciais no custo computacional e nos requisitos de memória.

O modelo *Transformer* processa as sequências de entrada, seguido por uma camada de *Mean Pooling* que consolida as representações em um vetor único. O *fine-tuning* é regido por estratégias de aprendizagem métrica[1], utilizando as funções de perda *Contrastive Loss* ou *Triplet Loss*. Tal abordagem assegura que os *embeddings* gerados na saída posicionem instâncias contextualmente similares em regiões próximas do espaço de representação, otimizando a discriminação entre as classes.

No que se refere ao pré-processamento, as *stopwords* foram preservadas, visto que a arquitetura BERT demonstra eficácia na extração de nuances contextuais a partir desses elementos. Por fim, o treinamento foi estabelecido com um limite de 100 épocas, utilizando o otimizador *Adam* com taxa de aprendizado de 10^{-4} e *batch size* de 16. Para mitigar o *overfitting* e assegurar a capacidade de generalização dos modelos, aplicou-se a técnica de *Early Stopping* com paciência de 5 ciclos, monitorando-se a convergência da função de perda no conjunto de validação.

3.2.1. Mecanismos de Aproximação e Distanciamento

Nesta abordagem, o codificador (DistilRoBERTa) ajusta os pesos de suas camadas para otimizar a qualidade dos *embeddings* via *Contrastive Loss* e *Triplet Loss*. O objetivo é o aprendizado de representações vetoriais onde instâncias semanticamente similares convirjam no espaço representação, enquanto exemplos dissimilaridades sejam repelidos.

A *Contrastive Loss* é aplicada utilizando a distância Euclidiana sobre pares de exemplos, conforme definido na Equação 1:

$$L = \frac{1}{2}(1-y)D^2 + \frac{1}{2}y\{\max(0, m-D)\}^2 \tag{1}$$

Em que y representa o rótulo binário (0 para similar, 1 para dissimilar), D denota a distância entre as representações e m é a margem de separação.

Complementarmente, a *Triplet Loss* utiliza triplas compostas por uma âncora (a), um exemplo positivo (p) e um negativo (n). O objetivo, expresso na Equação 2, assegura que a distância entre a âncora e o positivo seja inferior à distância entre a âncora e o negativo por uma margem m:

$$L = \max(0, D(a,p) - D(a,n) + m) \tag{2}$$

Para otimizar o aprendizado, empregou-se o *mining* de negativos *semi-hard*. Esses exemplos, que satisfazem a condição $D(a,p) < D(a,n) + m$, fornecem gradientes mais informativos e mitigam o *overfitting* em comparação a negativos *hard* [Kertész 2021]. Esse processo refina a capacidade discriminatória do modelo, permitindo que os *embed-*

[1]Aprendizagem métrica (ou *metric learning*) refere-se ao uso de algoritmos para aprender uma função de distância que capture a similaridade entre dados.

dings capturem relações semânticas profundas, como a ideologia de uma notícia, independentemente da fonte de publicação.

3.3. Tarefa de Classificação: Modelos e Parametrização

Após o mapeamento dos *embeddings*, em que a proximidade entre vetores reflete, além da similaridade semântica, a similaridade ideológica das notícias, a classificação dos artigos foi realizada por meio de três algoritmos: *K-Nearest Neighbors (KNN)*, *K-Means* e *Multilayer Perceptron (MLP)*. O *KNN* e o *K-Means* foram utilizados para explorar a organização dos dados por vizinhança e agrupamento, respectivamente. Para o *KNN*, aplicou-se *grid search* sistemático para otimização de hiperparâmetros, variando o número de vizinhos (k) em $\{5, 10, 15, 20, 25, 30\}$ e a função de distância entre as métricas euclidiana e cosseno. O *K-Means* foi configurado com número de *clusters* equivalente ao número de classes do conjunto ABP.

A rede *MLP* foi estruturada com duas camadas densas (512 e 256 neurônios), função de ativação *ReLU* nas camadas ocultas e *softmax* na camada de saída para classificação multiclasse. O treinamento da *MLP* foi realizado com otimizador *Adam*, função de perda *Categorical Cross-Entropy*, taxa de aprendizado de 10^{-3} e limite máximo de 100 épocas, configurações consolidadas na literatura para esse tipo de problema [Zhang and Rao 2020, Goodfellow et al. 2016].

Para todos os modelos avaliados, adotou-se validação cruzada com $k = 5$ *folds*. Esse procedimento mantém a proporção das classes nas partições, reduz vieses de avaliação e fornece uma estimativa mais robusta da capacidade de generalização em dados não vistos [Brink et al. 2016], sendo a média dos desempenhos obtidos nos *folds* utilizada como métrica final para a seleção do melhor modelo.

3.4. Avaliação de Desempenho

O desempenho dos modelos de classificação foi avaliado pelas métricas de Acurácia e *Macro F1-score*. A Acurácia (Equação 3) fornece uma medida geral da taxa de acerto para o conjunto de classes C. Complementarmente, o *Macro F1-score* (Equação 4) permite uma avaliação equilibrada entre as classes, mitigando distorções causadas por eventual desbalanceamento no conjunto de dados.

As métricas são formalmente definidas conforme segue:

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3) \qquad \text{Macro } F1 = \frac{1}{|C|} \sum_{c \in C} F1_c \quad (4)$$

Em que $F1_c$ representa a média harmônica entre a Precisão (P_c) e a Revocação (R_c) para cada classe:

$$P_c = \frac{TP_c}{TP_c + FP_c}, \quad R_c = \frac{TP_c}{TP_c + FN_c}, \quad F1_c = 2 \times \frac{P_c \times R_c}{P_c + R_c} \quad (5)$$

Neste contexto, TP, TN, FP e FN representam, respectivamente, os verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

A validação da hipótese de pesquisa — de que o discurso textual reflete o viés ideológico — dar-se-á mediante a obtenção de altos índices em ambas as métricas. Espera-se que valores elevados de *Macro F1-score* confirmem a capacidade discriminatória do modelo entre as diferentes vertentes ideológicas, assegurando que o desempenho não seja reflexo de uma classe majoritária no conjunto de dados.

3.5. Ambiente de Execução

A linguagem Python, com as bibliotecas NumPy, Pandas, Scikit-Learn e PyTorch, foi a ferramenta primária para implementação e avaliação dos modelos. Os experimentos ocorreram em um servidor equipado com processador Intel Xeon W-2235, 128 GB de RAM e GPU NVIDIA RTX 8000 (48 GB VRAM), visando a aceleração em hardware. Com o objetivo de favorecer a transparência e a reprodutibilidade dos experimentos, o código-fonte, os hiperparâmetros e os *scripts* de pré-processamento foram disponibilizados publicamente[2].

4. Resultados e Discussão

Para a tarefa de geração de *embeddings*, a análise do espaço de representação via t-SNE[3] na Figura 2 permite compreender como as diferentes funções de perda estruturam as representações do modelo DistilRoBERTa. O comparativo revela que a *Contrastive Loss* (Figura 2a) gera agrupamentos visivelmente mais densos e com fronteiras de separação nítidas entre as classes ideológicas. Em contrapartida, o mapeamento via *Triplet Loss* (Figura 2b) apresenta maior dispersão e zonas de sobreposição entre os *clusters*, o que compromete a precisão dos modelos de aprendizagem em cenários de maior complexidade.

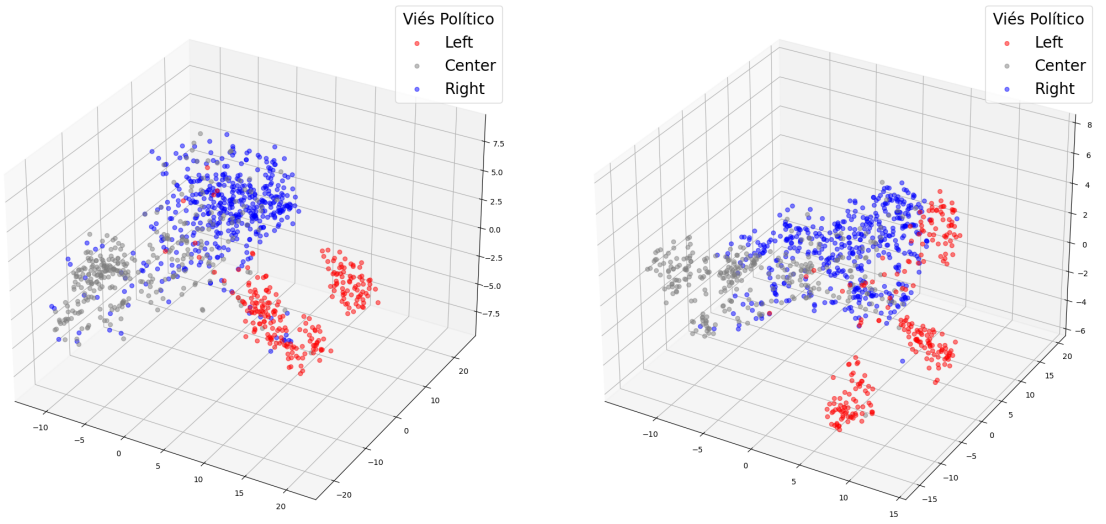
Além da classificação de polaridade ideológica, os *embeddings* aprendidos mostraram potencial de reutilização em tarefas analíticas complementares. Como essas representações preservam proximidade semântica entre textos, podem ser empregadas em recuperação de notícias similares, agrupamento por enquadramento discursivo e monitoramento temporal de mudanças de narrativa entre veículos. Na prática, esse reuso permite alimentar múltiplos módulos com a mesma base vetorial, reduzindo custo computacional e simplificando a operação em larga escala.

Em relação à tarefa de classificação de viés ideológico, cujos resultados encontram-se sintetizados nas Tabelas 4 e 5, as colunas associadas ao ABP reportam o desempenho interno na respectiva partição de teste do próprio conjunto ABP, enquanto as colunas associadas ao *FlipBias* reportam a avaliação externa do mesmo modelo em outro domínio. As métricas observadas no cenário *random split* mostram-se superiores àquelas obtidas via *media-bias split*, resultado compatível com a maior homogeneidade na distribuição dos dados de treinamento. Na partição *media-bias split*, o modelo fundamentado em MLP com *Contrastive Loss* atingiu o maior *Macro F1-score* $(0,43)$, indicando vantagem em relação à configuração com *Triplet Loss* nesse cenário.

No tocante à generalização para o conjunto *FlipBias*, o desempenho foi analisado principalmente pelo *Macro F1-score*. Considerando os modelos treinados no ABP com

[2]`https://github.com/jailsonpj/detecting-ideological-bias`

[3]Configuração do t-SNE: número de componentes=3, perplexidade=30, iterações=1000, métrica=euclidiana.



(a) *Contrastive Loss* com DistilRoBERTa. **(b) *Triplet Loss* com DistilRoBERTa.**

Figura 2. Distribuição dos *embeddings* via t-SNE no conjunto de teste do ABP. O comparativo demonstra a separação das classes entre as funções de perda *Contrastive* e *Triplet Loss*.

as funções de perda *Contrastive Loss* e *Triplet Loss*, o algoritmo *K-means* apresentou os melhores resultados relativos nas avaliações externas. Em particular, a combinação *Contrastive Loss* + *K-means* alcançou *Macro F1-score* de $0,41$ no *media-bias split* e $0,48$ no *random split*, indicando maior capacidade de transferência para o FlipBias em comparação às demais configurações desse recorte experimental.

4.1. Aplicações Complementares dos *Embeddings*

As representações vetoriais também viabilizam aplicações além da predição de classe. Em particular, as distâncias no espaço de *embeddings* podem apoiar auditoria editorial por similaridade entre matérias, organização de coleções por afinidade semântica e detecção de amostras fora de distribuição para priorização de rotulação humana. Esse cenário é especialmente relevante quando há necessidade de incorporar novos veículos ao sistema de monitoramento, pois o mesmo espaço vetorial pode sustentar classificação, busca semântica e curadoria de dados sem retrabalho integral do pipeline.

Apesar do desempenho geral competitivo, a análise por classe revela comportamento desigual entre os rótulos. No conjunto FlipBias, a melhor configuração experimental (AR-DR-C-KMEANS) apresentou baixa revocação para a classe *left*, atingindo apenas $0,31$. Esse resultado indica que parte significativa dos textos progressistas não foi corretamente identificada pelo modelo, sendo atribuída a outras classes. Uma possível explicação é que os exemplos *left* do FlipBias usem vocabulário, enquadramentos ou padrões discursivos diferentes daqueles observados no ABP. Portanto, mesmo com aprendizagem métrica, o modelo ainda permanece sensível ao domínio de treinamento, o que evidencia a dificuldade de transferir representações entre diferentes ecossistemas de notícias.

4.2. Desempenho de Generalização

A Tabela 6 demonstra que modelos treinados no FlipBias apresentam desempenho sólido no domínio de origem, com o classificador *KNN* atingindo $0,81$ de *Macro F1-Score*.

Tabela 4. Desempenho dos modelos treinados no conjunto ABP com *media-bias split*. As colunas ABP Media-bias indicam o desempenho na partição de teste do próprio ABP, enquanto as colunas FlipBias indicam a avaliação externa no conjunto FlipBias. C: Contrastive Loss; T: Triplet Loss; AM: ABP media split; DR: DistilRoBERTa; MF1: Macro F1-Score; Acc: Acurácia; Rec: Revocação; Prec: Precisão.

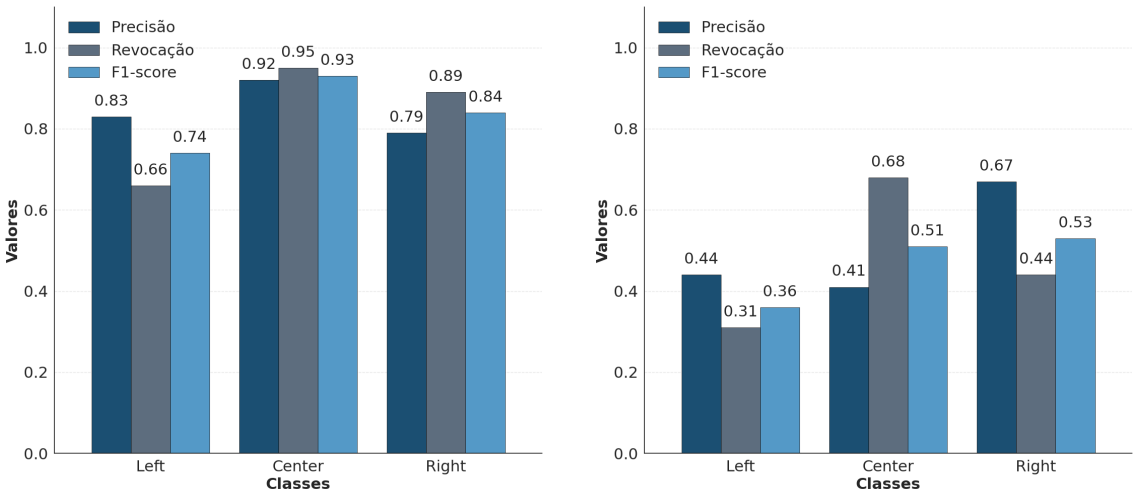
		ABP Media-bias				FLIPBIAS			
Perda	**Modelos**	**MF1**	**Acc**	**Rec**	**Prec**	**MF1**	**Acc**	**Rec**	**Prec**
Contrastive	AM-DR-C-KNN	0,31	0,31	0,36	0,36	0,24	0,28	0,28	0,30
	AM-DR-C-MLP	**0,43**	0,43	**0,49**	**0,47**	0,29	0,32	0,32	0,46
	AM-DR-C-kmeans	0,21	0,46	0,33	0,15	**0,41**	**0,43**	**0,43**	**0,56**
Triplet	AM-DR-T-KNN	0,23	0,25	0,30	0,31	0,34	0,42	0,42	0,38
	AM-DR-T-MLP	0,34	0,36	0,40	0,39	0,28	0,35	0,35	0,38
	AM-DR-T-kmeans	0,36	**0,50**	0,41	0,33	0,39	0,40	0,40	0,53

Tabela 5. Desempenho dos modelos treinados no conjunto ABP com *random split*. As colunas ABP Random indicam o desempenho na partição de teste do próprio ABP, enquanto as colunas FlipBias indicam a avaliação externa no conjunto FlipBias. C: Contrastive Loss; T: Triplet Loss; AR: ABP random split; DR: DistilRoBERTa; MF1: Macro F1-Score; Acc: Acurácia; Rec: Revocação; Prec: Precisão.

		ABP Random				FLIPBIAS			
Perda	**Modelos**	**MF1**	**Acc**	**Rec**	**Prec**	**MF1**	**Acc**	**Rec**	**Prec**
Contrastive	AR-DR-C-KNN	0,83	0,83	0,84	0,83	0,41	0,41	0,41	0,50
	AR-DR-C-MLP	0,80	0,80	0,82	0,79	0,42	0,42	0,42	0,48
	AR-DR-C-kmeans	**0,83**	**0,83**	0,83	**0,85**	**0,48**	0,47	**0,48**	**0,51**
Triplet	AR-DR-T-KNN	**0,83**	**0,83**	**0,85**	0,83	0,47	**0,48**	**0,48**	0,50
	AR-DR-T-MLP	0,83	0,83	0,84	0,83	0,43	0,45	0,45	0,50
	AR-DR-T-kmeans	0,79	0,79	0,82	0,78	0,38	0,48	0,48	0,37

Contudo, observa-se uma queda drástica de eficácia ao testar na base do ABP, onde o melhor resultado cai para $0,41$. Esse cenário indica que, embora a *Contrastive Loss* auxilie na extração de representações úteis, a generalização entre diferentes conjuntos de dados de viés midiático permanece um desafio crítico para arquiteturas baseadas em embeddings.

Ao comparar esses experimentos com os baselines na Tabela 7, nota-se que o modelo AR-DR-C-kmeans apresenta desempenho competitivo no conjunto ABP, alcançando $0,83$ de *Macro F1-Score* e acurácia. Esse resultado supera [Baly et al. 2020] no protocolo considerado, mas permanece abaixo do melhor resultado reportado por Volf e Simko [Volf and Simko 2025], de $0,84$ em ambas as métricas. Por outro lado, no conjunto FlipBias, o modelo F-DR-C-KNN destaca-se como o melhor classificador entre todos os comparados, atingindo $0,81$ de *Macro F1-Score*. Essa discrepância reforça que a escolha do conjunto de treinamento e a especialização do modelo ditam a confiabilidade da detecção em diferentes domínios.



(a) AR-DR-C-KMEANS no ABP. **(b) AR-DR-C-KMEANS no FlipBias.**

Figura 3. Resultados das métricas de precisão, acurácia e macro f1-score por classe para o modelo AR-DR-C-KMEANS avaliado nos conjuntos de dados ABP e FlipBias.

Tabela 6. Desempenho dos modelos treinados no conjunto FlipBias e avaliados no ABP. F: FlipBias; DR: DistilRoBERTa; C: Contrastive Loss; MF1: Macro F1-Score; ACC: Acurácia; Rec: Revocação; Prec: Precisão.

		FLIPBIAS				ABP			
Dados	**Modelos**	**MF1**	**Acc**	**Rec**	**Prec**	**F1**	**Acc**	**Rec**	**Prec**
	F-DR-C-KNN	**0,81**	**0,80**	**0,81**	**0,81**	**0,41**	**0,47**	**0,44**	**0,41**
FLIPBIAS	F-DR-C-MLP	0,81	0,80	0,81	0,81	0,26	0,32	0,37	0,24
	F-DR-C-kmeans	0,78	0,77	0,77	0,79	0,35	0,53	0,41	0,34

Por fim, a Tabela 8 mostra que LLMs em regime *zero-shot* podem apresentar desempenho competitivo, com destaque para a acurácia de $0,84$ do Mistral. Ainda assim, no protocolo adotado, os modelos especializados baseados em *embeddings* mantiveram melhor equilíbrio entre classes, refletido em *Macro F1-Score* superior. Assim, os resultados sugerem um compromisso prático: abordagens *decoder-only* oferecem boa capacidade de generalização sem ajuste supervisionado, enquanto pipelines leves com DistilRoBERTa e classificadores dedicados tendem a entregar desempenho comparável com menor custo de inferência quando treinados em dados aderentes ao domínio-alvo.

5. Considerações Finais

Este trabalho investigou a detecção de viés ideológico em notícias por meio de aprendizagem métrica profunda, combinando representações contextuais do DistilRoBERTa com funções de perda *Contrastive* e *Triplet*. Os resultados evidenciam que a configuração AR-DR-C-KMeans alcançou *Macro F1-score* de $0,83$ no conjunto ABP, com desempenho competitivo em relação aos principais baselines considerados.

Na avaliação cruzada entre domínios (ABP $\rightarrow$ FlipBias e FlipBias $\rightarrow$ ABP), observou-se queda de desempenho, indicando que a generalização para fontes inéditas

Tabela 7. Comparativo de desempenho entre os baselines e os experimentos nos conjuntos FLIPBIAS e ABP. MF1: Macro F1-Score, ACC: Acurácia.

	FLIPBIAS		ABP	
Modelos	**MF1**	**Acc**	**MF1**	**Acc**
[Baly et al. 2020]	-	-	0,80	0,79
AR-DR-C-kmeans	0,48	0,47	0,83	0,83
F-DR-C-KNN	**0,81**	0,80	0,41	0,47
[Lin et al. 2025]	0,51	**0,89**	0,63	0,62
[Volf and Simko 2025]	0,53	0,56	**0,84**	**0,84**

Tabela 8. Comparativo de desempenho entre os experimentos e LLMs no conjunto FLIPBIAS. MF1: Macro F1-Score, ACC: Acurácia.

Modelos	**Versão**	**Estratégia**	**MF1**	**Acc**
AR-DR-C-kmeans	-	-	0,48	0,47
F-DR-C-KNN	-	-	**0,81**	0,80
LLaMa2	Llama-2-7B-Chat	Zero-shot	0,52	0,69
Mistral	Mistral-7B-v0.1	Zero-shot	0,55	**0,84**
GPT-3.5	gpt-3.5-turbo	Zero-shot	0,39	0,15

permanece um desafio aberto. Esse achado reforça a importância de estratégias de treinamento orientadas à robustez de domínio, especialmente em tarefas sensíveis a variações discursivas e contextuais, e da inclusão de amostras representativas dos veículos-alvo.

Em comparação com LLMs em regime *zero-shot*, observou-se que modelos *decoder-only* podem alcançar resultados competitivos sem ajuste supervisionado. Por outro lado, os modelos especializados baseados em DistilRoBERTa apresentaram melhor equilíbrio entre classes no cenário avaliado, indicando que abordagens leves treinadas no domínio correto constituem alternativa prática para reduzir custo computacional em aplicações contínuas.

Adicionalmente, os resultados indicam que os *embeddings* treinados podem ser reaproveitados para outras finalidades além da classificação de polaridade, como recuperação semântica de notícias similares, agrupamento de narrativas e apoio à identificação de amostras candidatas à rotulação. Esse reuso amplia o valor prático da abordagem, pois permite atender múltiplas demandas analíticas com baixo custo de inferência e com a mesma infraestrutura representacional.

Como limitações, destacam-se: (i) a dependência de conjuntos de dados em língua inglesa e (ii) a sensibilidade a mudanças de domínio entre veículos. Como trabalhos futuros, pretende-se investigar adaptação de domínio, *few-shot prompting* e integração com *Retrieval-Augmented Generation*, além de mensurar latência e custo por amostra em cenários de uso contínuo.

Referências

Baly, R., Da San Martino, G., Glass, J., and Nakov, P. (2020). We can detect your bias: Predicting the political ideology of news articles. In *Proc. of EMNLP*, pages 4982–4991.

Brink, H., Richards, J., and Fetherolf, M. (2016). *Real-World Machine Learning*. Manning Publications.

Chen, W., Wachsmuth, H., Khatib, K., and Stein, B. (2018). Learning to flip the bias of news headlines. In *Proc. of INLG*, pages 79–88.

Chiang, C. and Knight, B. (2011). Media bias and influence: Evidence from newspaper endorsements. *The Review of Economic Studies*, 78:795–820.

Dallmann, A., Lemmerich, F., Zoller, D., and Hotho, A. (2015). Media bias in german online newspapers. In *Proc. of HT*, pages 133–137.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Efron, M. (2004). The liberal media and right-wing conspiracies. In *Proc. of ACM CIKM*, pages 390–398.

Gangula, R., Duggenpudi, S., and Mamidi, R. (2019). Detecting political bias in news articles using headline attention. In *Proc. of ACL BlackboxNLP*, pages 77–84.

Gentzkow, M. and Shapiro, J. M. (2006). Media bias and reputation. *Journal of Political Economy*, 114:280–316.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Kertész, G. (2021). Different triplet sampling techniques for lossless triplet loss on metric similarity learning. In *Proc. of IEEE SAMI*, pages 449–454.

Lin, L., Wang, L., Guo, J., and Wong, K. (2025). Investigating bias in llm-based bias detection: Disparities between llms and human perception. In *Proc. of COLING*, pages 10634–10649.

Lin, L., Wang, L., Zhao, X., Li, J., and Wong, K. (2024). Indivec: An exploration of leveraging llms for media bias detection. *arXiv preprint arXiv:2402.00345*.

Lin, Y., Bagrow, J., and Lazer, D. (2011). More voices than ever? quantifying media bias in networks. In *Proc. of ICWSM*, volume 5.

Liu, Y. et al. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y., Zhang, X., Wegsman, D., Beauchamp, N., and Wang, L. (2022). Politics: Pretraining with same-story article comparison for ideology prediction. In *Findings of NAACL*, pages 1354–1374.

Patricia Aires, V., Nakamura, F., and Nakamura, E. (2019). A link-based approach to detect media bias in news websites. In *Companion Proc. of WWW*, pages 742–745.

Spinde, T. et al. (2021). Neural media bias detection using distant supervision with babe. In *Findings of EMNLP*, pages 1166–1177.

Volf, M. and Simko, J. (2025). Political learning and politicalness classification of texts. *arXiv preprint 2507.13913*.

Zhang, Y. and Rao, Z. (2020). Deep neural networks with pre-train model bert for aspect-level sentiments classification. In *Proc. of IEEE ITOEC*, pages 923–927.