Instructions for Authors of SBC Conferences Papers and Abstracts

[1]

Abstract. *This meta-paper describes the style to be used in articles and short papers for SBC conferences. For papers in English, you should add just an abstract while for the papers in Portuguese, we also ask for an abstract in Portuguese ("resumo"). In both cases, abstracts should not have more than 10 lines and must be in the first page of the paper.*

Resumo. *Este meta-artigo descreve o estilo a ser usado na confecção de artigos e resumos de artigos para publicação nos anais das conferências organizadas pela SBC. É solicitada a escrita de resumo e abstract apenas para os artigos escritos em português. Artigos em inglês deverão apresentar apenas abstract. Nos dois casos, o autor deve tomar cuidado para que o resumo (e o abstract) não ultrapassem 10 linhas cada, sendo que ambos devem estar na primeira página do artigo.*

1. Introdução

2. Trabalhos Relacionados

3. Material e Métodos

A metodologia fundamenta-se na premissa de que o viés ideológico se manifesta em padrões discursivos e semânticos recorrentes, processando conteúdos textuais via modelos pré-treinados para geração de *embeddings* otimizados por *Contrastive* e *Triplet Loss*. Como ilustrado na Figura 1, o fluxo de trabalho compreende quatro etapas principais: a definição de conjuntos de treino e teste baseados na literatura, a geração de representações vetoriais via aprendizagem métrica, o treinamento de classificadores sobre vetores otimizados e a análise sistemática do desempenho. As seções que seguem descrevem detalhadamente cada módulo integrante da arquitetura apresentada.

3.1. Dados Experimentais

Para o desenvolvimento deste estudo, utilizou-se o conjunto de dados *Article Bias Prediction* (ABP), proposto por Baly $et\ al.$ [2020].O *corpus* compreende 30.246 artigos de notícias em língua inglesa, classificados de forma supervisionada nas categorias de viés ideológico: *left*, *center* e *right*. A confiabilidade dos rótulos advém da plataforma AllSides[1], que emprega um processo rigoroso de auditoria – incluindo revisões de terceiros e *feedback* da comunidade – garantindo que anotações reflitam com precisão a orientação política dos textos. A Figura 2 ilustra um exemplo de como um tópico de notícia foi classificado de acordo com a ideologia presente no texto.

[1] `https://www.allsides.com/media-bias/media-bias-rating-methods`

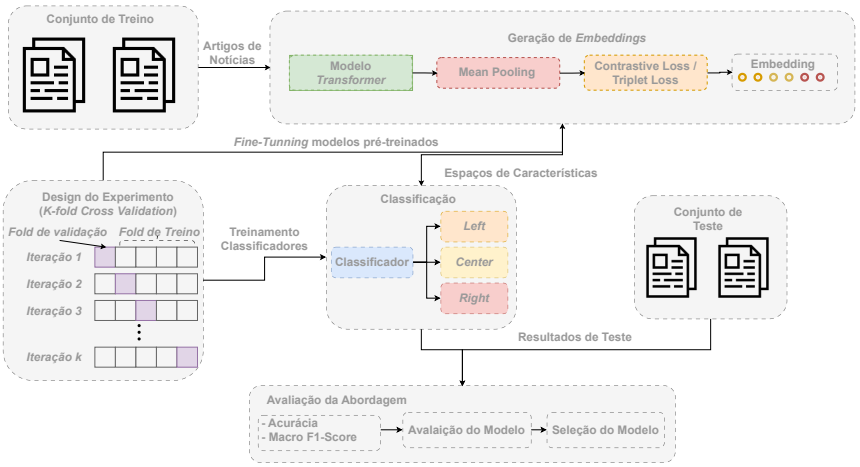

Figura 1. Visão geral do método de detecção de viés ideológico por meio do conteúdo textual de artigos de notícias.

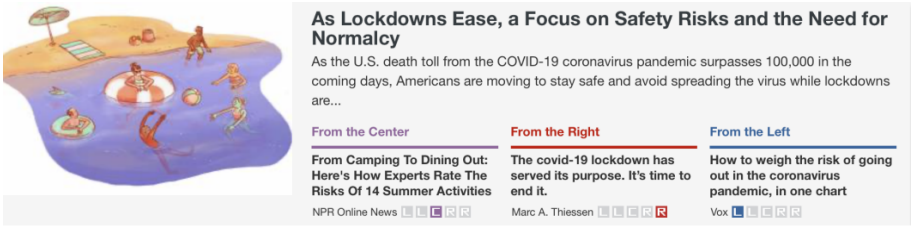

Figura 2. Exemplo de uma amostra de artigo de notícia com o tópico sobre a pandemia de coronavírus [Baly et al. 2020].

O ABP apresenta elevada diversidade temática, abrangendo desde processos eleitorais até questões sociais complexas. Visando assegurar que os modelos de aprendizado de máquina capturam a ideologia expressa linguisticamente, e não apenas identifiquem a fonte de notícia, os autores realizaram um pré-processamento para remover marcadores explícitos, como nomes de autores e de portais. Essa preocupação é fundamental para garantir a generalização do modelo e a integridade da análise discursiva, evitando que o classificador aprendizada vieses específicos de veículo de imprensa em vez de padrões semânticos.

A robustez da avaliação foi garantida através de dois métodos de particionamento: o *random split* e o *media-bias split*. Na Tabela 1, observa-se a configuração do *media-bias split*, no qual as fontes são segregadas para garantir que o modelo seja testado em veículos não expostos durante o treinamento, mitigando o vazamento de dados. Em contraste, o *random-split* permite a sobreposição de fontes entre as partições, mantendo a consistência na distribuição de classes entre treino e validação, como apresentado na Tabela 2. Neste trabalho, a totalidade das amostras do ABP foi empregada em todas as etapas metodológicas, desde a geração de *embeddings* até o treinamento e avaliação final.

Tabela 1. Estatísticas da partição *media-bias split*.

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 2. Estatísticas da partição *random split*.

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%

3.1.1. Avaliação de Generalização: Dataset FlipBias

Para avaliar o poder de generalização dos modelos experimentados, utilizou-se o conjunto de dados *FlipBias* [Chen et al. 2018]. Extraído da plataforma AllSides, o *corpus* organiza-se em 2.781 eventos noticiosos, cada qual coberto por perspectivas de diferentes inclinações políticas (*left*, *center* e *right*).

Para assegurar a comparabilidade dos resultados, os modelos foram submetidos às condições de avaliação estabelecidas por [Lin et al. 2024, Lin et al. 2025]. Essa abordagem permite verificar a robustez das representações na identificação de nuances ideológicas, garantindo que o desempenho observado reflita uma capacidade real de abstração frente a dados não utilizados durante o treinamento.

3.2. Tarefa de Geração de *Embeddings*

Para a execução da tarefa, a Figura 3 ilustra o fluxo de processamento adotado. O processo inicia-se com a *tokenização* dos artigos de notícias, seguida pela mineração de exemplos. Esta etapa é fundamental para selecionar amostras informativas que otimizam a convergência e o aprendizado do modelo.

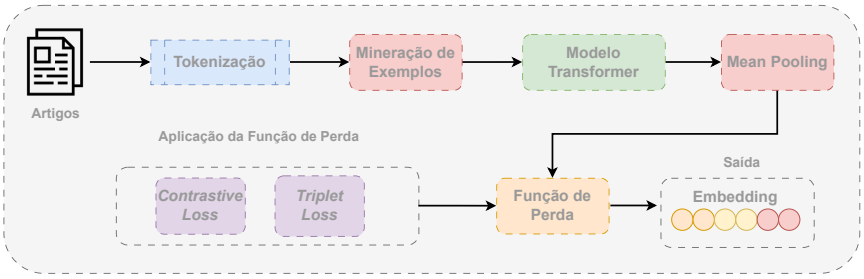

Figura 3. Fluxo de treinamento e geração de características para artigos de notícias.

Conforme delineado na arquitetura apresentada, empregaram-se dois modelos fundamentados em *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 DistilBERT [Sanh 2019] e o DistilRoBERTa [Liu 2019], variantes destiladas que preservam a robustez das arquiteturas originais, contudo, apresentam 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[2], 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 $0,0001$ 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 30 ciclos, monitorando-se a convergência da função de perda no conjunto de validação.

3.2.1. Mecanismos de Aproximação e Distanciamento

Nesta abordagem, os codificadores (DistilBERT e DistilRoBERTa) ajustam 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:

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

Onde 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

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

informativos e mitigam o *overfitting* em comparação a negativos *hard* [Kertész 2021]. Esse processo refina a capacidade discriminatória do modelo, permitindo que os *embeddings* capturem relações semânticas profundas, como a ideologia de uma notícia, independentemente da fonte de publicação.

3.3. Tarefa de Classificação: Modelos e Parametrizaçao

Após o mapeamento dos *embeddings*, onde a proximidade entre os veotres reflete 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 vizinhaça e agrupamento, respectivamente. Para o *KNN*, aplicou-se um *grid search* sistemático para otimização de hiperparâmetros, variando o número de vizinhos (k) entre $5, 10, 15, 20, 25$ e 30. Já o *K-Means* foi configurado com o número de *clusters* equivalente às classes presentes no conjunto de dados ABP.

A rede *MLP* foi estruturada com duas camadas densas (512 e 256 neurônios). Adotou-se a função de ativação *ReLU* para garantir um treinamento mais rápido e estável [Zhang and Rao 2020], enquanto a camada de saída utilizou a *softmax* para a classificação final. O modelo otimizado com o algoritmo *Adam* e a função de perda *Categorical Cross-Entropy*, escolhas consolidadas na literatura para problemas multiclasse [Goodfellow 2016].

A confiabilidade do experimento foi assegurada pela validação cruzada estritificada (5-fold). Esse procedimento garante que a proporção das classes seja mantida em todas as etapas, evitando resultados enviesados e permitindo medir com precisão a capacidade do modelo em classificar novos dados [Brink et al. 2016].

3.4. Avaliação de Desempenho

O desempenho dos modelos de classificação será 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 eventuais desbalanceamento no conjunto de dados.

As métricas são formalmente definidas conforme segue:

$$\text{Acurácia} = \frac{1}{|C|} \sum_{c \in C} \left(\frac{TP_c + TN_c}{TP_c + TN_c + FP_c + FN_c} \right) \tag{3}$$

$$\text{Macro } F1 = \frac{1}{C} \sum_{c \in C} F1_c \tag{4}$$

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

$$P_c = \frac{TP_c}{TP_c + FP_c}, \quad R_c = \frac{TP_c}{TP_c + FN_c} \tag{5}$$

$$F1_c = 2 \times \frac{P_c \times R_c}{P_c + R_c} \tag{6}$$

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. Comprometendo-se com a transparência e a reprodutibilidade técnica, o código-fonte, hiperparâmetros e *scripts* de pré-processamento estão publicamente disponíveis[3].

4. Resultados e Discussão

Os experimentos foram conduzidos seguindo a metodologia proposta, totalizando 24 configurações que integraram os modelos DistilBERT e DistilRoBERTa, funções de perda (*Contrastive Loss* e *Triplet Loss*), e classificadores como *KNN*, *K-Means* e *MLP*. Essa abordagem permitiu uma análise sistemática da sensibilidade dos modelos nos cenários *random split* e *media-bias split*, cujo os melhores desempenhos estão sintetizados na Tabela 3 em contraste com os *baselines* da literatura.

Tabela 3. Desempenho comparativo (em %) no conjunto de dados ABP utilizando as divisões *media-bias split* e *random split*.

Conj. de Dados	Modelo	Configuração (Repre. / Fun. de Perda / Classificador)	Macro F1-Score	Acurácia
ABP (media- bias split)	Proposto (Exp. 4)	DistilRoBERTa / Constrastive / KNN	45.44	48.89
	Baly el al., (2020)	BERT / - / BERT	33.53	36.75
ABP (random split)	Proposto (Exp. 16)	DistilRoBERTa / Constrastive / KNN	83.90	83.89
	Baly el al., (2020)	BERT / - / BERT	80.19	79.83

Os resultados demonstram que o uso da *Contrastive Loss* favorece a classificação de viés ideológico ao otimizar a separação vetorial no espaço de *embeddings*. Conforme a Tabela Tabela 3, o modelo proposto (*DistilRoBERTa + KNN*) atingiu $45,44\%$ de *Macro F1-score* no cenário *media-bias split*, superando a abordagem de Baly et al. [2020] em aproximadamente 12 pontos percentuais. No entanto, a análise dos particionamentos revela que o desempenho é significativamente superior no *random split* ($83,90\%$),

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

onde a presença de fontes comuns entre treino e teste facilita a predição. Já o *media-bias split* impõe um desafio de generalização maior, exigindo que o modelo identifique nuances ideológicas em portais inéditos, servindo como um estágio intermediário entre a memorização de fontes e a generalização plena.

Tabela 4. Desempenho comparativo (em %) no conjunto de dados FlipBias.

Modelo	Configuração (Repre. / Fun. de Perda / Classificador)	Estratégia	Macro F1-Score
Lin et al. (2024)	- / - / BERT	Fine-tuning	86.20
Lin et al. (2025)	- / - / GPT 3.5	Fine-tuning	77.82
Proposto (Exp. 16)	DistilRoBERTa / Constrastive / KNN	Fine-tuning	47.00
Proposto (Exp. 4)	DistilRoBERTa / Constrastive / KNN	Fine-tuning	28.21

Para testar os limites dessa generalização em um cenário de estresse ainda mais rigoroso, a Tabela 4 apresenta os resultados de transferência de domínio no conjunto *FlipBias*. Diferente dos *baselines* de Lin et al. [2024, 2025], que foram treinados e testados integralmente no *FlipBias*, as configurações propostas foram treinadas no *dataset* ABP e avaliadas sem exposição prévia ao novo domínio. Embora os experimentos 16 e 4 tenham obtido *Macro F1-Scores* de $47,00\%$ e $28,21\%$ — valores inferiores aos modelos *in-domain*[4] —, os dados evidenciam os desafios intrínsecos da avaliação *cross-domain*[5]. Essa análise valida a arquitetura ao demonstrar que, mesmo sem ajuste fino no domínio específico, o modelo captura estruturas ideológicas fundamentais, embora a especialização no domínio de destino continue sendo um fator determinante para o desempenho superior.

5. Considerações Finais

Este estudo investigou o impacto da aprendizagem métrica na classificação de viés político, revelando que a *Contrastive Loss* integrada ao *DistilRoBERTa* oferece uma base robusta para a separação semântica de discursos. Os resultados no conjunto ABP demonstram que a arquitetura proposta supera *baselines* consolidados mesmo no desafiador cenário *media-bias split*, mantendo uma capacidade de discernimento superior à literatura ao lidar com veículos de mídia inéditos. Evidenciou-se que, embora o desempenho seja otimizado pela proximidade entre treino e teste, o modelo captura estruturas ideológicas fundamentais que transcendem a simples memorização de fontes.

Por outro lado, os experimentos no *FlipBias* revelaram as fronteiras da generalização *cross-domain*, indicando que particularidades editoriais ainda influenciam os resultados em cenários sem ajuste fino. Para mitigar essa disparidade, trabalhos futuros focarão em técnicas de *Domain Adaptation*. Além disso, planeja-se expandir os modelos para análises multilíngues e testar modelos de *Large Language Models* (*LLMs*).

[4]In-domain refere-se a uma situação em que os dados utilizados para treinar um modelo de aprendizagem de máquina e os dados usados para testá-lo vêm da mesma distribuição, contexto ou fonte.

[5]Cross-domain refere-se ao treinamento de um modelo de aprendizagem de máquina em um determinado contexto e avaliado em um contexto diferente.

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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991.

Brink, H., Richards, J., and Fetherolf, M. (2016). Real-world machine learning. manning publications.

Chen, W.-F., Wachsmuth, H., Al Khatib, K., and Stein, B. (2018). Learning to flip the bias of news headlines. In *Proceedings of the 11th International conference on natural language generation*, pages 79–88.

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

Goodfellow, I. (2016). Deep learning.

Kertész, G. (2021). Different triplet sampling techniques for lossless triplet loss on metric similarity learning. In *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 000449–000454.

Lin, L., Wang, L., Guo, J., and Wong, K.-F. (2025). Investigating bias in LLM-based bias detection: Disparities between LLMs and human perception. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10634–10649, Abu Dhabi, UAE. Association for Computational Linguistics.

Lin, L., Wang, L., Zhao, X., Li, J., and Wong, K.-F. (2024). Indivec: An exploration of leveraging large language models for media bias detection with fine-grained bias indicators.

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

Sanh, V. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Zhang, Y. and Rao, Z. (2020). Deep neural networks with pre-train model bert for aspect-level sentiments classification. In *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)*, pages 923–927. IEEE.