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

1

Abstract. *This study investigates ideological bias identification using metric learning with BERT-based models. It demonstrates that Contrastive and Triplet Loss optimize embedding extraction for political classification, outperforming the literature baseline by 12 percentage points in rigorous evaluations. The model proved competitive against zero-shot LLMs, establishing itself as a robust, efficient alternative that captures semantic nuances beyond outlet memorization, ensuring integrity in analyzing unseen data.*

1. Introduction

The digitalization of news media has led to an unprecedented proliferation of online content, particularly through news portals. This growth presents the critical challenge of ensuring the integrity and impartiality of disseminated information, given that ideological bias can distort public perception and significantly influence public opinion [Gentzkow and Shapiro, 2006]. Previous studies have demonstrated that media bias affects how readers interpret political decisions and social discussions, potentially even influencing election outcomes [Chiang and Knight, 2011; DellaVigna and Kaplan, 2007]. However, identifying such bias remains a complex task due to the inherent subjectivity involved, making the development of automated detection methods a highly relevant and urgent area of research [Yigit-Sert et al., 2016].

The automatic detection of ideological bias represents a significant challenge within Natural Language Processing (NLP) and Machine Learning. Various studies have addressed this task by analyzing textual content [Krestel et al., 2012; Spinde et al., 2021], hyperlinks [Efron, 2004], and social media interactions [Ribeiro et al., 2018]. Despite these efforts, existing approaches often struggle with scalability and accuracy in broader contexts, as they are frequently limited to isolated metrics or specific case studies. Furthermore, many methodologies rely heavily on external metadata, which hinders the creation of autonomous solutions capable of identifying diverse types of bias across shifting contexts.

Recent literature has expanded this field by exploring Large Language Models (LLMs) and specialized pre-training objectives. For instance, the *POLITICS* framework introduces ideology-driven pre-training by comparing articles on the same story across different outlets [Liu et al., 2022], while tools like *IndiVec* leverage fine-grained indicators to improve adaptability [Lin et al., 2024]. However, these advancements introduce new complexities, such as inherent political biases within the LLMs themselves, which may manifest as preferences for specific viewpoints [Rozado, 2023] or create disparities between model predictions and human perception [Lin et al., 2025]. These findings underscore the necessity of developing more robust and transparent methodologies that

can navigate ideological nuances without being compromised by the biases of underlying architectures.

To mitigate these shortcomings, this work introduces a novel framework based on Deep Metric Learning to create optimized feature spaces that minimize the distance between news articles sharing the same ideological orientation. Unlike approaches limited to sentence-level classification, this research explores multiple textual aspects, including full-text content and probabilistic topic distribution.

By integrating these elements, this study aims to advance the state of the art through the following contributions:

- Deep Metric Learning Framework: The application of *Triplet Loss* and *Contrastive Loss* to generate optimized embedding spaces for full-text documents, specifically tailored for political bias detection.
- Multidimensional Ideological Analysis: A classification approach that encompasses a broader political spectrum, including centrist positions, moving beyond restrictive binary (left/right) models.
- Computational Efficiency and Accessibility: The implementation of a specialized architecture with a parameter count significantly lower than that of LLMs, facilitating high-performance inference on standard hardware without sacrificing accuracy.
- Autonomous and Source-Independent Classification: A methodology that operates exclusively on raw textual content, eliminating the dependency on external metadata or third-party knowledge bases.
- Empirical Performance Gains: Demonstration of superior effectiveness over traditional baselines by leveraging the synergy between Transformer architectures and metric learning techniques.

2. Related Work

The detection of ideological bias in news has been addressed through various lenses, ranging from structural networks to advanced linguistic modeling. Early systematic studies, such as those by Efron [2004], focused on structural elements like co-citation information to estimate political orientation. Similarly, Gentzkow and Shapiro [2006] proposed economic models to analyze how media reputations influence audience alignment. Advancing this structural view, Lin et al. [2011] combined social network analysis with NLP to quantify bias through interaction graphs between sources and entities. While effective, these methods often struggle with documents containing sparse hyperlink data or those isolated from established citation networks.

The analysis of textual content remains the most direct approach to bias detection. Dallmann et al. [2015] examined framing in German newspapers, identifying how word choice and theme frequency reflect ideological tendencies. Gangula et al. [2019] utilized attention mechanisms to capture bias specifically in headlines, highlighting their role in shaping reader perception. In a broader context, Baly et al. [2020] developed a robust framework using BERT and LSTM for textual classification. Their findings indicated that while high accuracy could be achieved on known sources, these models often failed to generalize when faced with news articles from unseen sources, a limitation that remains a central challenge in the field.

Social media platforms offer alternative metadata for bias inference through user interactions and demographic data. Rao and Spasojevic [2016] utilized word embeddings and LSTM networks to classify the political leaning of tweets, achieving high accuracy in binary U.S. political contexts. Elejalde et al. [2017] automated political orientation tests by analyzing news outlets' tweets and vocabulary. Furthermore, Ribeiro et al. [2018] inferred ideological orientation by analyzing ad interface data on Facebook and Twitter, estimating bias based on the demographic profile of a source's audience. Despite their effectiveness, these methods remain dependent on external platform data and are often restricted to highly polarized, country-specific scenarios.

The emergence of Large Language Models (LLMs) has introduced new paradigms, such as the *POLITICS* framework [Liu et al., 2022], which uses ideology-driven pre-training to compare articles across different media outlets. Similarly, *IndiVec* [Lin et al., 2024] leverages the instruction-following capabilities of LLMs combined with vector databases to provide fine-grained bias indicators. However, the use of LLMs introduces complexities, as studies by Rozado [2023] show that architectures like ChatGPT manifest consistent preferences for specific political viewpoints. Furthermore, significant disparities have been observed between LLM predictions and human perception of bias [Lin et al., 2025], necessitating more transparent and specialized architectures.

To address these gaps, this research explores Deep Metric Learning to create structured feature spaces where ideological proximity is modeled by distance metrics. This study introduces a source-independent method focusing on the article's raw text, utilizing localized textual segments to extract source-agnostic signatures. By applying *Triplet/Contrastive Losses*, the approach clusters articles by ideological similarity more cohesively than traditional classification heads, capturing the core editorial framing while maintaining the efficiency of standard Transformer encoders. This shift to metric representation learning extracts source-agnostic signatures, overcoming generalization hurdles and providing a computationally sustainable alternative to LLMs without compromising detection accuracy.

2.1. Summary and Comparison

The methods reviewed indicate that most approaches are either limited to specific case studies, restricted to polarized binary classifications, or dependent on external metadata. Structural methods often fail with sparse data, while social media-based methods remain platform-dependent. Although LLM approaches are powerful, they introduce concerns regarding inherent model bias and high computational costs. Table 1 summarizes these works, contrasting traditional models with recent LLM-based state-of-the-art results to highlight the critical trade-off between computational cost and predictive performance.

In this context, the model by Baly et al. [2020] is adopted as the primary baseline for direct comparison, as it was specifically trained and evaluated on the Articles Bias Prediction (ABP) dataset. Although LLM-based frameworks like *POLITICS* report high metrics (82.2% average accuracy) using massive pre-training on the BIGNEWS dataset, Baly et al. [2020] provides more comparable experimental conditions to evaluate the impact of metric learning architectures. This choice allows for isolating the model's effectiveness in identifying ideological nuances without relying on massive scales or external knowledge bases.

Tabela 1. Summary of the related work including LLM-based approaches and datasets.

Author	Strategy	Dataset	Bias Classes	Performance
Efron [2004]	Co-citation/hyperlinks	Open Directory	Liberal/Conservative	77.5% (Accuracy)
Lin et al. [2011]	Social Network Analysis	OpenCongress	Democrat/Republican	-
Dallmann et al. [2015]	Similarity metrics	German federal parliament	German political party	-
Rao et al. [2016]	Word embeddings + LSTM	Twitter API	Democrat/Republican	87% (Accuracy)
Elejalde et al. [2017]	Rank difference	Twitter API	Liberal/Conservative	-
Ribeiro et al. [2018]	Audience statistics	Facebook API	Liberal/Conservative	-
Baly et al. [2020]	BERT + LSTM	ABP	Left/Center/Right	79.83% (Macro F1)
Liu et al. [2022] (POLITICS)	Ideology-driven Pre-training	BIGNEWS	Left/Center/Right	82.2% (Avg. Acc.)
Lin et al. [2024] (IndiVec)	LLM + Vector Database	FLIPBIAS	Left/Center/Right	81.3% (Accuracy)
Lin et al. [2025]	LLM (zero-shot)	FLIPBIAS	Left/Center/Right	39.4% (Macro F1)

3. Proposed Approaches

In this section, We present the proposed approaches for classifying ideological bias in news articles using deep metric models to generate richer embedding representations, based both on their textual content and on the combination of textual representations with their topic probability distribution.

3.1. Textual Content-Based Method

The methodology adopted in this study is based on the hypothesis that articles sharing the same ideological bias exhibit similar discursive forms, demonstrating ideological alignment in both structure and content. This approach focuses on the analysis of textual content in news articles, specifically examining discursive disparities between sources with distinct ideological orientations. To achieve this, pre-trained models are employed for the generation of textual embeddings, which are subsequently refined through fine-tuning using *Contrastive Loss* and *Triplet Loss* to effectively capture the nuances of political bias. As illustrated in Figure 1, the methodological process follows a sequential workflow consisting of data acquisition based on related works, the generation of news article embeddings optimized via the aforementioned loss functions, the training of a classification model utilizing these refined textual representations through machine learning techniques, and the final evaluation of the results.

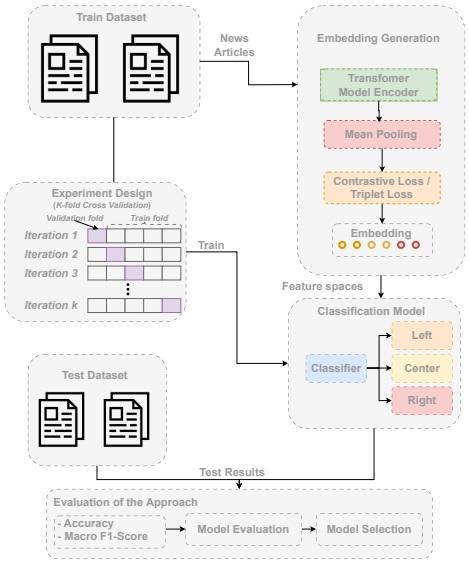

Figura 1. Overview of the textual content-based classification method.

3.1.1. Dataset

Data acquisition followed the methodology of Baly et al. [2020], utilizing the Article Bias Prediction (ABP) dataset. Articles were sourced from diverse news portals and labeled via the AllSides platform, which ensures high reliability through rigorous editorial audits and independent reviews.

The dataset comprises $30,246$ English-language news articles, supervisedly classified into three categories: left, right, and center. This robust labeling facilitates precise analysis of discursive divergences, providing a solid foundation for training and testing machine learning models across various political contexts.

Tabela 2. Statistics of the Article Bias Prediction (ABP) dataset.

Ideological Bias	Quantity	Percentage (%)
Left	12,003	34.55
Center	9,743	28.05
Right	12,991	37.40
Total	**34,737**	**100.00**

As detailed in Table 2, the ABP statistics show a distribution of 34.6% left, 37.3% right, and 28.1% center, covering diverse topics from elections to social issues. This study utilizes the full dataset for embedding generation, topic extraction, and final classification model evaluation.

3.1.2. Embedding Generation

This research utilizes pre-trained models based on Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2018] to capture long-range semantic dependencies and complex contextual relationships. This architecture enables a deep understanding of sentence structure, resulting in more effective vector representations for natural language processing tasks [Gao et al., 2021].

The study evaluates DistilBERT [Sanh, 2019], a compact version focused on computational efficiency, and DistilRoBERTa [Liu, 2019], which offers superior performance due to optimized large-scale training. Both models underwent fine-tuning via Metric Learning, applying *Triplet Loss* and *Contrastive Loss* functions to ensure that semantically similar examples occupy adjacent spaces within the embedding domain.

Data processing involved specific tokenizers standardized to 512 tokens, as illustrated in Figure 3. Stopwords were retained to preserve the contextual nuances inherent in BERT-based models. Finally, Mean Pooling was applied across the token dimension to generate a single vector representation that synthesizes the most relevant features of each document.

3.2. Loss Functions

In this approach, the DistilBERT and DistilRoBERTa encoders adjust their layer weights to enhance embedding quality through *Contrastive* and *Triplet Loss* functions. The objective is to produce vector representations where similar ideological examples converge in the embedding space while dissimilar ones diverge. Consequently, the encoders are optimized to cluster articles by ideological alignment rather than source-specific characteristics.

Contrastive Loss is implemented using Euclidean distance (D) between pairs of positive and negative examples, as defined in Equation 1:

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

where y is the binary similarity label and m is the margin. Conversely, the model with *Triplet Loss* utilizes triplets composed of an anchor (a), a positive example (p), and a negative example (n), following Equation 2:

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

to ensure that the positive example remains closer to the anchor than its negative counterpart.

To refine discriminative performance, the methodology incorporates the mining of semi-hard negative samples, identifying triplets where $D(a, p) < D(a, n) + m$. This technique provides more informative gradients and mitigates *overfitting* compared to hard negative samples [**?**]. As illustrated in Figure 4, this process ensures that the model clusters articles by ideology—such as matching an anchor with a positive case from a different news outlet—thereby generating source-independent semantic signatures.

Referências