# PRISM: A Framework for Producing Interpretable Political Bias Embeddings with Political-Aware Cross-Encoder

**Yiqun Sun[1], Qiang Huang[2]*, Anthony K. H. Tung[1], Jun Yu[2]**
[1]School of Computing, National University of Singapore
[2]School of Intelligence Science and Engineering, Harbin Institute of Technology (Shenzhen)
{sunyq, atung}@comp.nus.edu.sg, {huangqiang, yujun}@hit.edu.cn

## Abstract

Semantic Text Embedding is a fundamental NLP task that encodes textual content into vector representations, where proximity in the embedding space reflects semantic similarity. While existing embedding models excel at capturing general meaning, they often overlook ideological nuances, limiting their effectiveness in tasks that require an understanding of political bias. To address this gap, we introduce **PRISM**, the first framework designed to **P**roduce inte**R**pretable pol**I**tical bia**S** e**M**beddings. PRISM operates in two key stages: (1) Controversial Topic Bias Indicator Mining, which systematically extracts fine-grained political topics and their corresponding bias indicators from weakly labeled news data, and (2) Cross-Encoder Political Bias Embedding, which assigns structured bias scores to news articles based on their alignment with these indicators. This approach ensures that embeddings are explicitly tied to bias-revealing dimensions, enhancing both interpretability and predictive power. Through extensive experiments on two large-scale datasets, we demonstrate that PRISM outperforms state-of-the-art text embedding models in political bias classification while offering highly interpretable representations that facilitate diversified retrieval and ideological analysis. The source code is available at `https://github.com/dukesun99/ACL-PRISM`.

## 1 Introduction

Semantic Text Embedding is a fundamental NLP task that encodes texts into vector representations, where proximity in the embedding space reflects semantic similarity. This task has garnered extensive research attention (Mikolov et al., 2013; Pennington et al., 2014; Devlin et al., 2019; Gao et al., 2021; Zhuo et al., 2023; Li and Li, 2024) due to its broad applications in retrieval (Karpukhin et al., 2020; Thakur et al., 2021), clustering (Aggarwal and Zhai, 2012), and semantic textual similarity (STS) (Agirre et al., 2012, 2013).

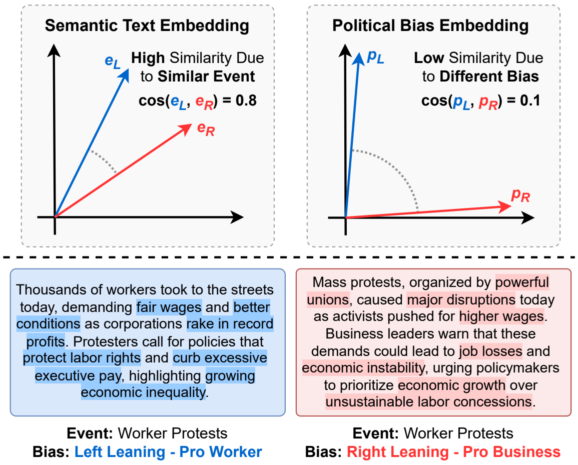


Figure 1: Semantic Text Embedding vs. Political Bias Embedding: While the former captures event-level similarity, the latter reveals ideological orientation in news coverage.

While significant progress has been made in developing advanced models for semantic representation, these approaches primarily capture general meaning while often failing to account for underlying political bias. This limitation becomes apparent in real-world scenarios such as news coverage of the same event–illustrated in Figure 1–where articles may report on identical topics but convey distinctly different political perspectives. Existing embedding models (Gao et al., 2021; Zhuo et al., 2023; Li and Li, 2024), despite assigning high similarity scores based on shared content, struggle to capture these ideological nuances, exposing a crucial gap in current methodologies.

To address this gap, this work investigates **Political Bias Embedding**, a new representation learning task that transforms textual content into compact vector representations customized to capture ideological orientations. These specialized embeddings offer significant advantages for various down-

*Qiang Huang is the corresponding author.



*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27719–27733
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

stream applications, ranging from political bias classification (Iyyer et al., 2014; Liu et al., 2022) to diversified retrieval systems that expose content across the ideological spectrum (Sun et al., 2024; Huang et al., 2024).

Existing research on political bias analysis has evolved from simple left-right classification (Iyyer et al., 2014; Baly et al., 2020a) to more nuanced approaches that consider multiple finer-grained political dimensions (Kim and Johnson, 2022; Liu et al., 2023). In parallel, advances in text embeddings have progressed from general-purpose representations (Reimers and Gurevych, 2019) to domain-specific models tailored for specialized fields (Lee et al., 2020; Beltagy et al., 2019). Specifically, for political text analysis, recent work has developed specialized encoder-only Transformer models (Devlin et al., 2019; Liu et al., 2019) through fine-tuning with political news articles (Liu et al., 2022), while instruction-following text embedding models have emerged as another promising direction (Su et al., 2023; Peng et al., 2024).

Despite these advancements, developing effective political bias embeddings remains challenging.

- **Complexity of Political Dimensions:** While political bias analysis has moved beyond binary classification toward multi-dimensional perspectives (Kim and Johnson, 2022; Liu et al., 2023), enumerating a comprehensive political bias taxonomy remains challenging. Automatically extracting fine-grained political topics from a news corpus is essential for learning robust and interpretable representations of political bias.
- **Scarcity of Fine-Grained Annotations:** High-quality, large-scale datasets capturing nuanced political perspectives are scarce, as manually annotating political bias is resource-intensive and subjective (Sinno et al., 2022; Kim and Johnson, 2022; Liu et al., 2023).
- **Black-box Models:** While recent studies have explored interpretable embeddings using yes/no questions as dimensions (McInerney et al., 2023; Benara et al., 2024; Sun et al., 2025), these approaches are not directly applicable to political bias analysis due to the complexity and subjectivity of political viewpoints. To the best of our knowledge, no existing work has systematically addressed the challenge of constructing interpretable embeddings for political bias.

To overcome these challenges, we introduce **PRISM**, the first framework developed to **P**roduce inte**R**pretable pol**I**tical bia**S** e**M**bedding. PRISM comprises two core stages:

- **Controversial Topic Bias Indicator Mining:** We extract fine-grained political topics and bias indicators from weakly labeled news data, addressing the scarcity of fine-grained annotations through an automated topic discovery process.
- **Cross-Encoder Political Bias Embedding:** We use mined topics as interpretable dimensions and bias indicators as reference points for bias scoring, developing a weak-label training strategy to enable nuanced political bias comprehension with a political-aware cross-encoder while designing topic retrieval to enhance efficiency.

**Contributions.** The key contributions of this work are summarized as follows:

(1) **Task Formulation and Framework Design:** We introduce the novel task of political bias embedding and present PRISM, the first framework specifically designed to generate interpretable political bias embeddings that go beyond surface-level semantics to capture ideological orientation.

(2) **Annotation-Free, Interpretable Embedding Approach:** PRISM operates without requiring fine-grained manual annotations. It leverages distant supervision to automatically mine controversial political topics and their associated bias indicators. Utilizing a political-aware cross-encoder, PRISM produces embeddings that are inherently interpretable, with dimensions explicitly grounded in semantically meaningful bias indicators.

(3) **Empirical Effectiveness Across Tasks:** Extensive experiments show that PRISM consistently outperforms state-of-the-art baselines on key downstream tasks such as political bias classification and politically diversified content retrieval, while offering transparent insight into ideological representation.

## 2 Related Work

**Political Bias Mining.** Political bias in news articles has been extensively studied (Baly et al., 2020b; Nakov et al., 2024; Martinez et al., 2024), with much of the research focused on political bias classification. Early work predominantly framed this as a binary classification problem, distinguishing between left-leaning and right-leaning viewpoints (Iyyer et al., 2014; Chen et al., 2017; Kulkarni et al., 2018; Fan et al., 2019; Baly et al., 2020a; Spinde et al., 2021; Kim and Johnson, 2022; Liu et al., 2022, 2023; Hong et al., 2023; Lin et al., 2024; Liu et al., 2024).

SLAP4SLIP by Hofmann et al. (2022) focuses on concept discovery and framing analysis, modeling ideological polarization along the dimensions of salience and framing. It employs graph neural networks with structured sparsity to detect polarized concepts without relying on explicit political orientation labels. While related, their objective differs from ours, as SLAP4SLIP emphasizes framing analysis rather than producing interpretable political bias embeddings. Media framing analysis, as surveyed by Otmakhova et al. (2024), offers another lens for understanding political bias by examining how information is "packaged" to elicit specific interpretations, often through emphasis or word choice. Complementary to this, recent work by Sutter et al. (2024) shows that integrating text embeddings with network structures via graph neural networks enhances performance in unsupervised stance detection.

Yet, recognizing the limitations of a single left-to-right scale, particularly in non-U.S. political contexts, researchers have explored multi-dimensional methods that classify texts by ideological stances on specific policy issues (e.g., gun control, abortion) (Kim and Johnson, 2022) or broader ideological dimensions like economic equality and political regime (Liu et al., 2023, 2024).

Beyond classification, politically diversified retrieval systems have been developed to surface content from across the ideological spectrum (Wu et al., 2020; Draws et al., 2021; Vrijenhoek et al., 2021; Huang et al., 2024; Sun et al., 2024). However, existing approaches largely rely on categorical labels or metadata rather than embedding-based representations of political bias, limiting their ability to generalize across diverse sources and viewpoints.

**Domain-Specific Embedding.** Domain-specific embedding models have proved effective in specialized fields such as biomedical (Chen et al., 2019; Lee et al., 2020), financial (Anderson et al., 2024; Tang and Yang, 2024), and scientific (Beltagy et al., 2019) domains. These models leverage domain-adaptive pretraining to better capture nuances within their respective fields.

For political bias analysis, POLITICS (Liu et al., 2022) fine-tunes RoBERTa (Liu et al., 2019) with bias-specific objectives, while recent instruction-following models (Su et al., 2023; Peng et al., 2024) enable the creation of task-specific embeddings. Despite their effectiveness, these models lack interpretability, making it difficult to understand and explain how ideological biases are embedded in their representations (Martinez et al., 2024).

**Interpretable Text Embedding.** Recent work on interpretable text embeddings either focuses on analyzing existing embeddings (Lee et al., 2022; Simhi and Markovitch, 2023) or generating inherently interpretable ones (Opitz and Frank, 2022; McInerney et al., 2023; Patel et al., 2023). A notable strategy for achieving interpretability using question-answer pairs as embedding dimensions, providing explicit semantic meaning for each dimension (Benara et al., 2024; Sun et al., 2025).

However, to our knowledge, no existing work addresses the unique challenges of creating interpretable embeddings for political bias analysis. Given the complexity and subjectivity of ideological viewpoints, existing frameworks lack mechanisms to explicitly encode and explain political bias. Our work bridges this gap by introducing the first framework, PRISM, to produce interpretable political bias embeddings, ensuring both effectiveness and transparency in ideological representation.

## 3 The PRISM Framework

### 3.1 Overview

We introduce PRISM, a framework designed to generate interpretable political bias embeddings for news articles. As illustrated in Figure 2, PRISM operates in two key stages:

(1) **Controversial Topic Bias Indicator Mining:** Extracts fine-grained political topics and their corresponding bias indicators from a large, weakly labeled news corpus (Section 3.2).

(2) **Cross-Encoder Political Bias Embedding:** Generates structured embeddings that encode political bias along interpretable dimensions (Section 3.3).

**Mining Political Topics and Bias Indicators.** PRISM first identifies controversial topics and their bias indicators from weakly labeled data. For example, a topic extracted from the BigNews dataset (Liu et al., 2022), "*Medicaid Overhaul and Health Care Funding,*" has two bias indicators:

- **Left Indicator:** "*Advocates for increased funding for public health, express concern over privatization and potential reductions in services.*"
- **Right Indicator:** "*Focus on cost control, support for block grants and private insurance options, prioritize reducing government spending.*"

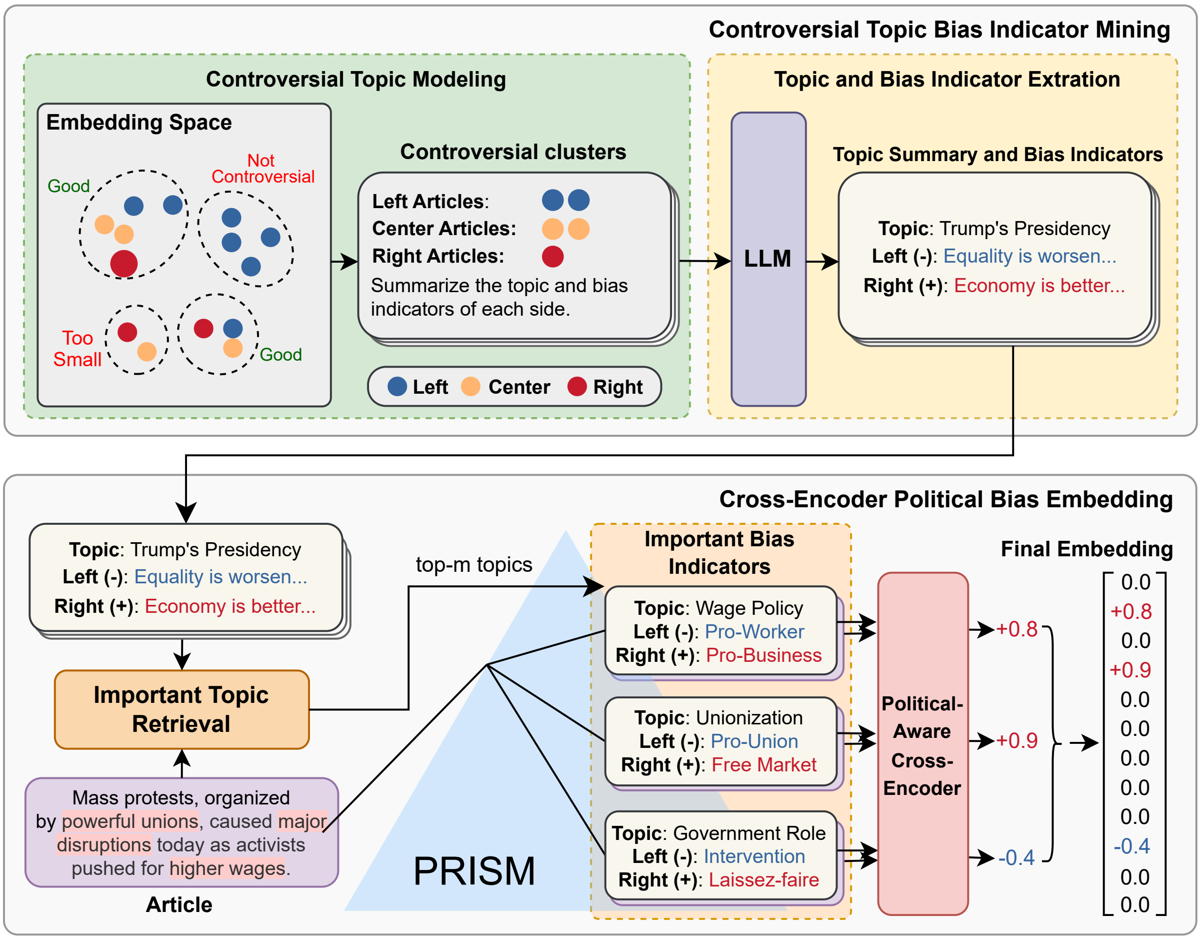


Figure 2: Overview of the PRISM framework.

These topics form the embedding dimensions of political bias, while the bias indicators serve as reference points for encoding political bias.

**Generating Interpretable Bias Embeddings.** To quantify an article's stance on each topic, PRISM employs a political-aware cross-encoder model, which assigns a score between 0 and 1 based on how strongly the article aligns with a given bias indicator. The resulting embeddings satisfy two essential interpretability properties:

- **Selective Activation:** Only relevant and bias-bearing topics receive nonzero values.
- **Explicit Bias Representation:** High scores indicate strong alignment with a specific bias, facilitating direct interpretation.

Within these two stages, PRISM enhances interpretability while maintaining flexibility across diverse ideological landscapes. We begin by detailing the controversial topic bias indicator mining stage.

### 3.2 Controversial Topic Bias Indicator Mining

**Motivation.** Capturing political bias requires fine-grained embedding dimensions that reflect ideological differences. However, manually curating such dimensions is costly and impractical. PRISM addresses this challenge by *automatically identifying controversial topics* and their associated *bias indicators* from a weakly labeled news corpus.

**Weakly Labeled News Corpus.** PRISM relies on weakly labeled media bias ratings rather than manually annotated data. Since news outlets often exhibit editorial biases through selective reporting or omission of certain facts (Baly et al., 2020a; Rodrigo-Ginés et al., 2024), PRISM utilizes datasets such as NewsSpectrum (Sun et al., 2024) and BigNews (Liu et al., 2022) as our weakly labeled news corpus, where articles are assigned *media bias ratings* (e.g., $-1$ for left, 0 for center, 1 for right) based on AllSides,[1] which provides expert-based bias assessments.

**Controversial Topic Modeling.** PRISM identifies controversial topics by leveraging semantic text encoders and clustering techniques:

(1) **Encoding the News Corpus:** Each article is transformed into an embedding using a pre-

[1] https://www.allsides.com/media-bias/ratings

trained semantic text encoder, which maps semantically similar texts to nearby locations in the embedding space.

(2) **Clustering Similar Articles:** Using $k$-means clustering, articles covering similar topics are grouped together.

(3) **Measuring Bias Dispersion:** The Bias Dispersion metric quantifies ideological diversity within each cluster by computing the variance of media bias ratings. Clusters with high dispersion indicate controversial topics, as they contain articles from diverse ideological views. Formally, for a cluster containing $n$ articles with bias ratings $\boldsymbol{R} = \{r_1, r_2, \cdots, r_n\}$, Bias Dispersion is calculated as:

$$\text{Bias Dispersion}(\boldsymbol{R}) = \frac{1}{n}\sum_{i=1}^{n}(r_i - \bar{r})^2,$$

where $\bar{r}$ is the mean bias rating of the cluster. Clusters with a Bias Dispersion exceeding a threshold $\tau$ and containing at least $p$ articles are identified as controversial topics, ensuring they are widely debated and well-represented.

**Topic and Bias Indicator Extraction.** For each identified topic, PRISM extracts bias indicators using LLMs. Specifically, given a set of sample articles and their media bias ratings, the LLM generates: (1) a concise, neutral topic summary and (2) bias indicators describing left-leaning and right-leaning perspectives. This process allows PRISM to systematically mine topics and define bias indicators without requiring *manual annotation*. The LLM prompt is provided in Appendix A.

### 3.3 Cross-Encoder Political Bias Embedding

To generate interpretable political bias embeddings, PRISM assigns values to each controversial topic dimension for a given article. This process consists of two pivotal steps: (1) Important Topic Retrieval, which identifies the most bias-bearing and relevant topics for the given news article; (2) Political-Aware Cross-Encoder Embedding, which computes bias alignment scores to generate structured embeddings.

**Important Topic Retrieval.** A key challenge in bias representation is ensuring that only *relevant* and *bias-bearing* dimensions are assigned nonzero values while filtering out *neutral* or *irrelevant* topics. To achieve this, PRISM retrieves the most important topics using pre-trained embeddings of the topics and their corresponding bias indicators.

For each news article, we encode its text, along with the topic summaries and their left and right indicators, using the same semantic text encoder from the previous topic mining stage. We then compute an importance score for each topic $i$ using the following equation:

$$\text{Score}(i) = \lambda(\boldsymbol{x} \cdot \boldsymbol{t}_i) + (1-\lambda)|\boldsymbol{x} \cdot \boldsymbol{r}_i - \boldsymbol{x} \cdot \boldsymbol{l}_i|, \quad (1)$$

where $\boldsymbol{x}$ is the embedding of the news article; $\boldsymbol{t}_i$ is the embedding of topic $i$; $\boldsymbol{l}_i$ and $\boldsymbol{r}_i$ are the embeddings of the left and right indicators of topic $i$; and $\lambda$ is the weighting factor balancing topic relevance and bias divergence.

Equation 1 balances two core factors: (1) *Relevance to the topic* $(\boldsymbol{x} \cdot \boldsymbol{t}_i)$ that measures how closely the article aligns with the topic; (2) *Bias divergence* $|\boldsymbol{r}_i \cdot \boldsymbol{x} - \boldsymbol{l}_i \cdot \boldsymbol{x}|$, which captures how strongly the article leans toward one bias over the other. By selecting the top-$m$ topics with the highest scores, PRISM ensures that embeddings remain efficient and interpretable, focusing only on the most relevant and bias-revealing topics.

**Political-Aware Cross-Encoder Embedding.** To ensure that the final embedding selectively activates only for relevant bias dimensions, we develop a new political-aware cross-encoder model that explicitly compares articles with bias indicators rather than relying solely on textual features.

**Training the Cross-Encoder Model.** A cross-encoder model (Nogueira and Cho, 2019), parameterized by $\theta$, takes two text inputs and outputs a bias alignment score:

$$f_\theta(\boldsymbol{a}, \boldsymbol{b}) \in (0, 1), \quad (2)$$

where $\boldsymbol{a}$ denotes a news article; $\boldsymbol{b}$ represents the bias indicator (left or right). Equation 2 quantifies the alignment between $\boldsymbol{a}$ and $\boldsymbol{b}$.

**Weak Label Generation for Training.** To train a political-aware cross-encoder, we generate weak supervision labels by leveraging the bias indicators of each topic cluster. Given an article $\boldsymbol{a}$, we first consider its *in-cluster*'s left $\boldsymbol{b}^{\text{left}}$ and right $\boldsymbol{b}^{\text{right}}$ indicators and create the following training pairs:

$$(\boldsymbol{a}, \boldsymbol{b}^{\text{left}}) \rightarrow \begin{cases} 1 & \text{if } bias = \text{left} \\ 0 & \text{otherwise} \end{cases},$$

$$(\boldsymbol{a}, \boldsymbol{b}^{\text{right}}) \rightarrow \begin{cases} 1 & \text{if } bias = \text{right} \\ 0 & \text{otherwise} \end{cases}.$$

Additionally, we consider some random *out-of-cluster* topics as negative samples, ensuring that the model does not falsely associate an article with unrelated topics:

$$(\boldsymbol{a}, \boldsymbol{b}^{\text{left}}) \to 0, \ (\boldsymbol{a}, \boldsymbol{b}^{\text{right}}) \to 0.$$

The model is trained using Mean Squared Error (MSE) loss. By learning to map article–indicator pairs to these labels, it is expected that the model can accurately distinguish relevant biases from neutral content and focus on topic-specific bias signals rather than generic political leanings.

**Generating the Final Bias Embedding.** During inference, PRISM produces the final embedding vector by computing bias alignment scores for the top-$m$ important topics.

Specifically, given a trained cross-encoder $f_\theta$, we generate the bias embedding as follows. First, we initialize the embedding vector with a list of zero values. Then, we retrieve the top $m$ most important topics $\boldsymbol{M}$ for the article $\boldsymbol{a}$. Lastly, we compute bias alignment scores using the cross-encoder for each selected topic. For each topic $i \in \boldsymbol{M}$, the final embedding value $e_i$ is computed as:

$$e_i = \begin{cases} s_i^r - s_i^l, & \text{if } i \in \boldsymbol{M}, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

where $s_i^l = f_\theta(\boldsymbol{a}, \boldsymbol{b}_i^{\text{left}})$ and $s_i^r = f_\theta(\boldsymbol{a}, \boldsymbol{b}_i^{\text{right}})$ are the bias alignment scores with the left indicator $\boldsymbol{b}_i^{\text{left}}$ and the right indicator $\boldsymbol{b}_i^{\text{right}}$.

**Remarks.** In Equation 3, PRISM encodes political bias by computing the difference between left and right bias scores, ensuring: (1) Positive values indicate a right-leaning bias; (2) Negative values indicate a left-leaning bias; and (3) Zero values indicate neutrality or irrelevance. This cross-encoder embedding approach guarantees that PRISM's embeddings remain interpretable, efficient, and explicitly tied to bias-revealing dimensions, validated through empirical analysis in Section 4.

# 4 Experiments

We evaluate the effectiveness of PRISM by addressing the following key research questions:

- **RQ1 (Political Bias Signal Quality):** How well does PRISM capture bias-related signals compared to generic semantic text embeddings and political bias-specific models? (Section 4.3)
- **RQ2 (Distance Measurement Effectiveness):** To what extent does the political bias embedding space accurately reflect ideological similarities between articles, making it suitable for diversified retrieval? (Section 4.4)
- **RQ3 (Interpretability):** Can the political bias embedding provide meaningful insights that users can intuitively interpret? (Section 4.5)

Before presenting results, we outline the datasets and benchmark models used for evaluation.

## 4.1 Datasets

We conduct experiments on two large-scale, real-world news datasets that provide extensive political coverage and diverse ideological perspectives.

- **NewsSpectrum** (Sun et al., 2024) consists of 250,000 news articles sourced from 961 distinct media outlets, with each article assigned a bias score ranging from left (-2) to right (2). This dataset is carefully curated to maintain a balanced distribution of political perspectives, making it particularly useful for evaluating models in diverse ideological settings.
- **BigNews** (Liu et al., 2022) comprises 3.6 million news articles collected from 13 media outlets, each labeled with its corresponding media bias. This large-scale dataset provides broad coverage of political discourse across various events and ideological stances.

## 4.2 Benchmark Models

Since PRISM is the first framework designed to produce interpretable political bias embeddings, there is no direct competitor. To systematically assess its performance, we compare it against state-of-the-art models from four relevant areas:

- **Generic Semantic Text Embedding Models:** We select **AnglE** (`UAE-Large-V1`) (Li and Li, 2024), a state-of-the-art text embedding model widely used for various NLP tasks (Muennighoff et al., 2023), as a strong baseline for generic semantic embeddings.
- **Political Bias-Specific Models:** We include **POLITICS** (Liu et al., 2022), the leading model for political bias analysis, pre-trained on BigNews. For evaluation, we use the official Hugging Face checkpoint `launch/POLITICS` and extract embeddings from the `CLS` token's last hidden state.
- **Instruction-Following Embedding Models:** We evaluate two cutting-edge instruction-following embedding models: **InstructOR** (`instructor-large`) (Su et al., 2023) and **InBedder** (`roberta-large-InBedder`) (Peng et al., 2024), provided with specific instructions for political bias analysis.

- **Interpretable Text Embedding Models:** To evaluate interpretability, we compare against **CQG-MBQA** (Sun et al., 2025), the state-of-the-art interpretable text embedding model. We use its publicly pre-trained checkpoint.[2]

### 4.3 Political Bias Classification

To assess how effectively PRISM captures political bias in news articles, we evaluate its performance on a political bias classification task.

**Experimental Setup.** We train an SVM classifier using embedding vectors generated by different models. Training is performed on a held-out dataset, distinct from PRISM's training data, and performance is evaluated on a separate test set. The classification results are presented in Table 1.

**Result Analysis.** As shown in Table 1, PRISM outperforms all baseline models, achieving the highest classification accuracy on NewsSpectrum and the second highest on BigNews. These results demonstrate PRISM's ability to effectively capture political bias signals while maintaining interpretability. Unlike generic semantic embeddings such as AnglE, which primarily encode overall content similarity, PRISM explicitly models ideological orientations, leading to superior performance.

Notably, while POLITICS achieves higher accuracy than PRISM on BigNews, this advantage stems from the test set of BigNews being part of its training data. Yet, when evaluated on NewsSpectrum, which was not seen during training, POLITICS lags significantly behind PRISM. This suggests that PRISM generalizes better across different datasets, reinforcing its ability to capture ideological bias without overfitting to specific corpora.

Overall, this experiment demonstrates that PRISM not only produces interpretable embeddings but also retains strong predictive power, establishing it as a robust and generalizable framework for political bias analysis.

### 4.4 Diversified Retrieval

A fundamental characteristic of political bias embeddings is their ability to serve as distance metrics for ideological similarity, making them particularly valuable for retrieval tasks. To evaluate this capability, we conduct politically diversified retrieval experiments, following the DiversiNews framework (Sun et al., 2024).

[2]https://github.com/dukesun99/CQG-MBQA

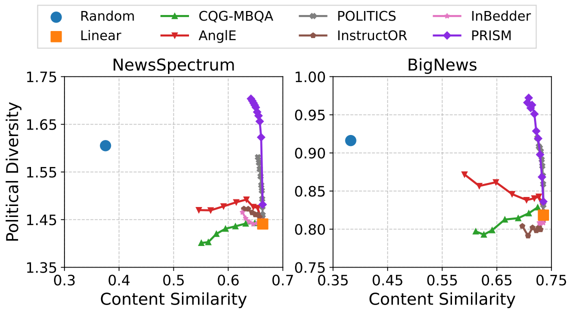


Figure 3: Diversified retrieval results.

**Experimental Setup.** We adopt the retrieval protocol and evaluation metrics from DiversiNews (Sun et al., 2024) and employ the Diversity-aware $k$-Maximum Inner Product Search (D$k$MIPS) algorithm (Huang et al., 2024) to enhance political diversity in retrieval results. Our implementation incorporates two distinct embedding spaces:

- **AnglE Embeddings:** Used to measure query-document relevance based on inner product similarity.
- **Model-specific Embeddings** (e.g., PRISM, POLITICS, InstructOR): Used to quantify inter-document political diversity based on ideological differences.

This dual-space design enables us to systematically evaluate each model's capacity to encode and differentiate political bias in retrieval scenarios.

We evaluate retrieval performance using the following measures:

- **Content Similarity:** Given a retrieved set $\boldsymbol{S} = \{\boldsymbol{p}1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_k\}$ that contains $k$ news articles for a query $\boldsymbol{q}$, the content similarity is defined as the mean inner product between each retrieved article $\boldsymbol{p}_i$ and the query $\boldsymbol{q}$:

$$\mathrm{Sim}(\boldsymbol{S}, \boldsymbol{q}) = \frac{1}{k} \sum_{i=1}^{k} \langle \boldsymbol{p}_i, \boldsymbol{q} \rangle,$$

where $\langle \boldsymbol{p}_i, \boldsymbol{q} \rangle$ represents the inner product similarity between article $\boldsymbol{p}_i$ and query $\boldsymbol{q}$. Higher values indicate stronger content relevance.

- **Political Diversity:** To measure the political diversity of retrieved results, we compute the mean pairwise difference between the bias ratings $r_i$ of articles in $\boldsymbol{S}$:

$$\mathrm{Div}(\boldsymbol{S}) = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} |r_i - r_j|.$$

Higher values indicate a greater spread of political perspectives within the retrieved articles $\boldsymbol{S}$, ensuring ideological balance in the results.

| Model | NewsSepctrum | | | | | BigNews | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc ↑ | Pre ↑ | Rec ↑ | F1-Ma ↑ | F1-Mi ↑ | Acc ↑ | Pre ↑ | Rec ↑ | F1-Ma ↑ | F1-Mi ↑ |
| AnglE | 48.4 | 48.2 | 48.8 | 48.2 | 48.4 | 69.0 | 69.0 | 69.0 | 69.0 | 69.0 |
| Instructor | 47.9 | 47.8 | 48.5 | 47.7 | 48.0 | 63.7 | 63.7 | 63.7 | 63.7 | 63.7 |
| InBedder | 50.2 | 50.2 | 50.8 | 49.8 | 50.2 | 64.6 | 64.6 | 64.6 | 64.6 | 64.6 |
| CQG-MBQA | 45.1 | 44.9 | 45.5 | 44.9 | 45.1 | 61.0 | 61.0 | 61.0 | 61.0 | 61.0 |
| POLITICS | 51.3 | 51.7 | 51.9 | 51.1 | 51.3 | **85.7** | **85.7** | **85.7** | **85.7** | **85.7** |
| **PRISM** | **86.1** | **86.5** | **86.2** | **86.2** | **86.1** | 73.5 | 81.3 | 73.7 | 74.0 | 73.5 |

Table 1: Political bias classification results on NewsSpectrum and BigNews, evaluated using Accuracy (Acc), Precision-Macro (Pre), Recall-Macro (Rec), F1-Macro (F1-Ma), and F1-Micro (F1-Mi).

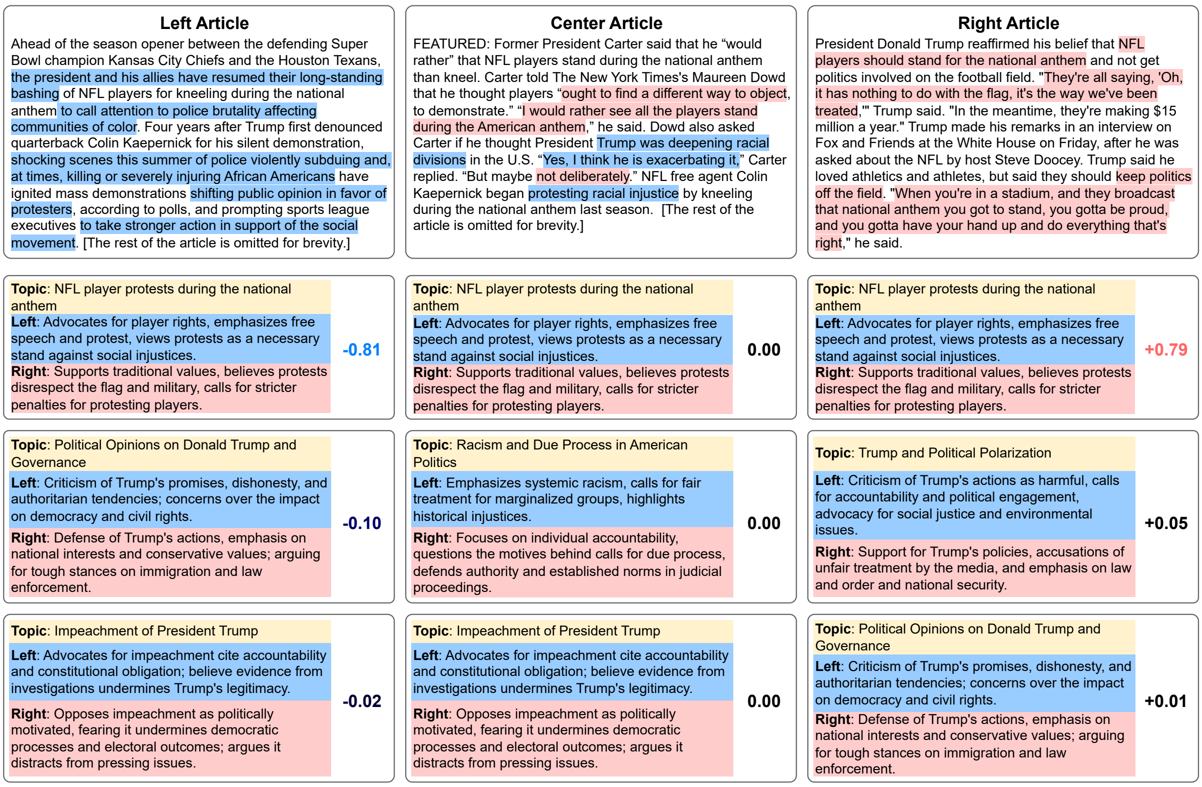


Figure 4: Case study on three news articles with different political bias.

**Result Analysis.** Figure 3 illustrates the trade-off between content similarity and political diversity across different retrieval methods. The results show that PRISM consistently outperforms all baselines in both dimensions: (1) At equivalent levels of political diversity, PRISM preserves higher content relevance than competing models; (2) At comparable content similarity, it delivers greater ideological diversity in the retrieved results.

This consistent advantage across both datasets highlights two key strengths of PRISM:

- **Better Bias Representation:** PRISM's embedding space more effectively captures ideological relationships between news articles compared to existing approaches.
- **Reliable Distance Metric:** The embeddings produced by PRISM serve as an effective metric for politically diversified retrieval, balancing relevance and ideological diversity.

## 4.5 Case Study

To illustrate PRISM's interpretability, we present a detailed case study analyzing its embedding representations for three news articles with different political orientations (Figure 4).

**Overview.** The selected articles examine the “**NFL player protests during the national anthem (2015-2020)**,” a politically charged topic with clear ideological divides. Each article is sourced from media outlets with distinct political leanings, reflecting contrasting bias indicators:

- **Left-leaning:** Highlights player rights, freedom

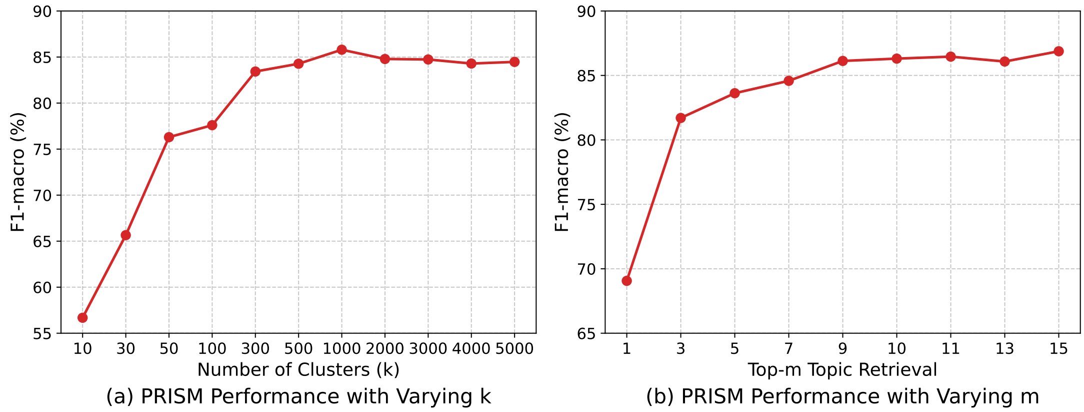


Figure 5: Effect of key parameters on classification performance (F1-marco) on NewsSpectrum: (a) number of clusters $k$ for topic mining; (b) number of top-$m$ topic dimensions retrieved per article.

of expression, and social justice.

- **Right-leaning:** Focuses on traditional values, patriotism, and respect for national symbols.

To aid visual interpretation, opinion-bearing text is color-coded: blue for left-leaning indicators and red for right-leaning ones.

**Findings and Interpretability Analysis.** PRISM exhibits several key capabilities that enhance its interpretability and efficacy in bias representation:

- **Accurate Topic Identification:** PRISM accurately identifies NFL player protests as the primary topic and assigns bias scores aligned with each article's ideological stance (-1 for left, 0 for center, and 1 for right).
- **Nuanced Topic Weighting:** Related topics receive proportionally smaller values, reflecting their secondary relevance. Unrelated topics are assigned values close to zero, ensuring embedding sparsity and interpretability.
- **Clear and Intuitive Bias Representation:** Users can interpret embeddings via bias indicators, while sparsity ensures only relevant dimensions hold meaningful values, minimizing noise and enhancing clarity.

### 4.6 Parameter Study

We conduct a parameter study to analyze the sensitivity effects of two key components in PRISM: (1) the number of clusters $k$ used in topic mining and (2) the number of top-$m$ topic dimensions retrieved per article during classification. All experiments are conducted on the NewsSpectrum dataset, with results summarized in Figure 5.

**Effect of Number of Clusters ($k$).** We vary the number of clusters $k$ used in the $k$-means topic mining stage from 10 to 5,000, without re-training the cross-encoder model. To ensure fair comparison, we proportionally adjust the minimum cluster size ($p$). As shown in Figure 5(a), performance improves steadily as $k$ increases, peaking around $k = 1{,}000$, after which it begins to decline slightly. This highlights the importance of choosing an appropriate number of clusters: too few clusters limit topic diversity, while too many may introduce noise or redundancy in the topic space.

**Effect of Top-$m$ Topic Retrieval ($m$).** We also evaluate the influence of the number of top-$m$ topic dimensions retrieved for each article. As illustrated in Figure 5(b), increasing $m$ leads to improved classification performance, with the F1-macro score rising steadily from $m = 1$ to $m = 9$, and stabilizing beyond that. This suggests that retrieving multiple relevant ideological dimensions enriches the representation, while further expansion offers diminishing returns.

## 5 Conclusions

In this work, we introduce Political Bias Embedding, a new task aimed at representing ideological orientations in a structured and interpretable manner. We propose PRISM, the first framework designed to produce interpretable political bias embeddings. By integrating automated topic mining with a new political-aware cross-encoder embedding approach, PRISM effectively captures political bias while maintaining interpretability. Extensive experiments demonstrate PRISM's superiority over existing models in political bias classification and diversified retrieval. Unlike standard semantic embeddings, PRISM encodes ideological distinctions while offering transparent bias insights, making it ideal for bias-aware retrieval and analysis.

## Limitations

Despite PRISM's strengths in robustness and interpretability for political bias embedding, several limitations remain:

**Efficiency and Scalability.** Although PRISM delivers strong performance, its two-stage design–consisting of topic mining and cross-encoder-based embedding–incurs higher computational cost compared to standard embedding models. This trade-off is justified by its substantial gains in interpretability and predictive accuracy. Nevertheless, future work could explore more efficient alternatives, such as model distillation and lightweight encoders, to improve scalability and reduce inference latency without sacrificing performance.

**Topic Granularity and Representation Assumptions.** PRISM relies on clustering-based topic extraction, which necessitates careful tuning of the number of clusters. Too few clusters risk oversimplifying ideological nuance, while too many may introduce noise or lead to overly sparse embeddings. Moreover, PRISM treats each topic dimension as an independent axis in the embedding space, implicitly assuming orthogonality among topics. While the use of top-$m$ topic retrieval allows for soft activation across multiple dimensions, this representation may still overlook inter-topic dependencies and correlated ideological dimensions. Future work could explore more expressive embedding structures that capture semantic overlap or hierarchical relationships among topics.

Furthermore, the current framework may conflate topic and stance, as fine-grained clusters often reflect both thematic content and ideological framing. Explicit disentanglement of these two elements, potentially through multi-view representation learning or factorized embeddings, could enhance the interpretability and generalizability of the learned representations.

**Evaluation Across Languages, Cultures, and Time.** Our current evaluation is limited to English-language news from U.S.-based media, annotated using AllSides bias ratings. Extending PRISM to other linguistic and cultural contexts would require region-specific corpora and localized bias references, as ideological dimensions may vary significantly across geographies. While PRISM is designed to be extensible, which enables new political topics to be incorporated via re-running the topic mining process on updated corpora, we do not explicitly evaluate its temporal robustness or responsiveness to emerging discourse. Future work may investigate time-sensitive evaluations to assess how well PRISM adapts to shifting ideological landscapes, including the emergence of new topics or changes in framing over time.

## Acknowledgments

Yiqun Sun and Anthony T. H. Tung were supported by the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001), and by the National Research Foundation, Singapore, under its AI Singapore Programme (AISG Award No. AISG3-RP-2022-029). Qiang Huang and Jun Yu were supported by the National Natural Science Foundation of China under grant Nos. 62125201 and U24B20174. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore, or the National Research Foundation, Singapore.

## References

Charu C Aggarwal and ChengXiang Zhai. 2012. A survey of text clustering algorithms. *Mining Text Data*, pages 77–128.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43.

Peter Anderson, Mano Vikash Janardhanan, Jason He, Wei Cheng, and Charlie Flanagan. 2024. Greenback bears and fiscal hawks: Finance is a jungle and text embeddings must adapt. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 362–370.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020a. 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.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020b. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3364–3374.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

Vinamra Benara, Chandan Singh, John X Morris, Richard J Antonello, Ion Stoica, Alexander G Huth, and Jianfeng Gao. 2024. Crafting interpretable embeddings for language neuroscience by asking llms questions. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 124137–124162.

Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2019. BioSentVec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5.

Wei Chen, Xiao Zhang, Tengjiao Wang, Bishan Yang, and Yi Li. 2017. Opinion-aware knowledge graph for political ideology detection. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3647–3653.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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 (NAACL-HLT)*, pages 4171–4186.

Tim Draws, Nava Tintarev, and Ujwal Gadiraju. 2021. Assessing viewpoint diversity in search results using ranking fairness metrics. *ACM SIGKDD Explorations Newsletter*, 23(1):50–58.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6343–6349.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6894–6910.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Valentin Hofmann, Xiaowen Dong, Janet Pierrehumbert, and Hinrich Schütze. 2022. Modeling ideological salience and framing in polarized online groups with graph neural networks and structured sparsity. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 536–550.

Jiwoo Hong, Yejin Cho, Jiyoung Han, Jaemin Jung, and James Thorne. 2023. Disentangling structure and style: Political bias detection in news by inducing document hierarchy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5664–5686.

Qiang Huang, Yanhao Wang, Yiqun Sun, and Anthony KH Tung. 2024. Diversity-aware $k$-maximum inner product search revisited. *arXiv preprint arXiv:2402.13858*.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1113–1122.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Michelle YoungJin Kim and Kristen Johnson. 2022. CLoSE: Contrastive learning of subframe embeddings for political bias classification of news media. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pages 2780–2793.

Vivek Kulkarni, Junting Ye, Steven Skiena, and William Yang Wang. 2018. Multi-view models for political ideology detection of news articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3518–3527.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Seonghyeon Lee, Dongha Lee, Seongbo Jang, and Hwanjo Yu. 2022. Toward interpretable semantic textual similarity via optimal transport-based contrastive sentence learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5969–5979.

Xianming Li and Jing Li. 2024. AoE: Angle-optimized embeddings for semantic textual similarity. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1825–1839.

Luyang Lin, Lingzhi Wang, Xiaoyan Zhao, Jing Li, and Kam-Fai Wong. 2024. IndiVec: An Exploration of Leveraging Large Language Models for Media Bias Detection with Fine-Grained Bias Indicators. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1038–1050.

Songtao Liu, Ziling Luo, Minghua Xu, Lixiao Wei, Ziyao Wei, Han Yu, Wei Xiang, and Bang Wang. 2023. Ideology takes multiple looks: A high-quality dataset for multifaceted ideology detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4200–4213.

Songtao Liu, Bang Wang, Wei Xiang, Han Xu, and Minghua Xu. 2024. Encoding hierarchical schema via concept flow for multifaceted ideology detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2930–2942.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022. POLITICS: Pretraining with Same-story Article Comparison for Ideology Prediction and Stance Detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374.

Manuel Martinez, Sonja Schmer-Galunder, Zoey Liu, Sangpil Youm, Chathuri Jayaweera, and Bonnie Dorr. 2024. Balancing transparency and accuracy: A comparative analysis of rule-based and deep learning models in political bias classification. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pages 102–115.

Denis McInerney, Geoffrey Young, Jan-Willem van de Meent, and Byron C Wallace. 2023. Chill: Zero-shot custom interpretable feature extraction from clinical notes with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8477–8494.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, pages 3111–3119.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2014–2037.

Preslav Nakov, Jisun An, Haewoon Kwak, Muhammad Arslan Manzoor, Zain Mujahid, and Husrev Sencar. 2024. A survey on predicting the factuality and the bias of news media. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15947–15962.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

Juri Opitz and Anette Frank. 2022. SBERT studies Meaning Representations: Decomposing Sentence Embeddings into Explainable Semantic Features. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*, pages 625–638.

Julia Otmakhova, Shima Khanehzar, and Lea Frermann. 2024. Media Framing: A Typology and Survey of Computational Approaches Across Disciplines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 15407–15428.

Ajay Patel, Delip Rao, Ansh Kothary, Kathleen Mckeown, and Chris Callison-Burch. 2023. Learning interpretable style embeddings via prompting llms. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15270–15290.

Letian Peng, Yuwei Zhang, Zilong Wang, Jayanth Srinivasa, Gaowen Liu, Zihan Wang, and Jingbo Shang. 2024. Answer is all you need: Instruction-following text embedding via answering the question. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 459–477.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. 2024. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications (ESA)*, 237:121641.

Adi Simhi and Shaul Markovitch. 2023. Interpreting embedding spaces by conceptualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1704–1719.

Barea Sinno, Bernardo Oviedo, Katherine Atwell, Malihe Alikhani, and Junyi Jessy Li. 2022. Political ideology and polarization: A multi-dimensional approach. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 231–243.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural media bias detection using distant supervision with babe-bias annotations by experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121.

Yiqun Sun, Qiang Huang, Yixuan Tang, Anthony K. H. Tung, and Jun Yu. 2025. A general framework for producing interpretable semantic text embeddings. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Yiqun Sun, Qiang Huang, Yanhao Wang, and Anthony K. H. Tung. 2024. DiversiNews: Enriching News Consumption with Relevant Yet Diverse News Articles Retrieval. *Proceedings of the VLDB Endowment*, 17(12):4277–4280.

Maia Sutter, Antoine Gourru, Amine Trabelsi, and Christine Largeron. 2024. Unsupervised stance detection for social media discussions: A generic baseline. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1782–1792.

Yixuan Tang and Yi Yang. 2024. Do we need domain-specific embedding models? an empirical investigation. *arXiv preprint arXiv:2409.18511*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*.

Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. 2021. Recommenders with a mission: assessing diversity in news recommendations. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (CHIIR)*, pages 173–183.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2020. SentiRec: Sentiment Diversity-aware Neural News Recommendation. In *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing (AACL-IJCNLP)*, pages 44–53.

Wenjie Zhuo, Yifan Sun, Xiaohan Wang, Linchao Zhu, and Yi Yang. 2023. Whitenedcse: Whitening-based contrastive learning of sentence embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 12135–12148.

## A Prompts

The following prompt is used to mine controversial topics and their associated bias indicators from news articles. For each cluster of articles, we provide the article texts and their corresponding media bias labels. The prompt instructs the language model to:

(1) Identify a common topic that reflects the primary point of contention
(2) Summarize the topic neutrally
(3) Extract distinct bias indicators for both left and right political perspectives
(4) Output the extracted topics and bias indicators in a structured format

This prompt design ensures consistent, structured extraction of topics and their associated political perspectives while maintaining neutrality in topic descriptions.

> Please summarize the following texts into a common topic, which the VAST MAJORITY of the texts debate on, and can reflect the bias of the texts, which different biases (left, center, right) hold different views on this topic.
> Note that there are multiple sides to the topic. Please summarize the topic in a neutral tone. Please return the topic and the bias indicators, without any other words or sentences. Please summarize the topic in fewer than 10 words.
>
> Give me the results in the following format: Topic: ⟨Topic⟩
> Left Indicator: ⟨Some key points that Left or Lean Left have⟩
> Right Indicator: ⟨Some key points that Right or Lean Right have⟩
>
> Text: {News_Article_i}
> Bias: {Media_Bias_of_News_Article_i}
>
> {Remaining Texts and Biases Omitted}
> ...
> ...

## B Implementation Details

Table 2 summarizes the hyperparameters used in the experiments.

**Controversial Topic Bias Indicator Mining.** We use `UAE-Large-V1` as the pre-trained encoder and apply $k$-means clustering with $k = 3{,}000$ to identify topic clusters. During the cluster filtering step, we set dataset-specific thresholds: a Bias Dispersion threshold of $\tau = 1.0$ for NewsSpectrum and $\tau = 0.5$ for BigNews, with a minimum cluster size of $p = 50$ for both. These thresholds are calibrated to account for the different label granularities in each dataset. For topic summarization and bias indicator generation, we randomly sample 50 texts per prompt and use `GPT-4o-mini` as the language model. This process yields 1,810 topics for NewsSpectrum and 2,279 topics for BigNews.

**Cross-Encoder Political Bias Embedding.** In the important topic retrieval component, we set the weighting factor $\lambda = 0.8$ to balance topic relevance and bias divergence. The political-aware cross-encoder model is implemented using `microsoft/deberta-v3-large` (304M) (He et al., 2023), trained with weak labels using a learning rate $\alpha = 10^{-6}$ and batch size $b = 4$. Training continues until loss convergence, which occurs at approximately 1,900,000 steps.

**Political Bias Classification.** We maintain the original label taxonomies for both datasets: a five-point scale {-2, -1, 0, 1, 2} for NewsSpectrum and a three-point scale {-1, 0, 1} for BigNews. We randomly sample 10,000 articles from NewsSpectrum and 100,000 articles from BigNews. Using scikit-learn version 1.5.2, we train an SVM classifier with default parameters[3] on 90% of the evaluation partition and test on the remaining 10%. All results are reported from a single experimental run.

**Diversified Retrieval.** For the politically diversified retrieval experiment, we employ an extended version of the BC-Greedy-Avg algorithm of D$k$MIPS (Huang et al., 2024) that operates in dual embedding spaces. The objective function is formulated as:

$$f(\boldsymbol{S}) = \frac{\lambda}{k} \sum_{i \in \boldsymbol{S}} \langle \boldsymbol{p_i}, \boldsymbol{q} \rangle - \frac{2\mu(1-\lambda)}{k(k-1)} \sum_{i \neq j \in \boldsymbol{S}} \langle \hat{\boldsymbol{p}}_i, \hat{\boldsymbol{p}}_j \rangle,$$

where $\boldsymbol{S}$ denotes the result set, $\boldsymbol{q}$ represents the query vector, $\boldsymbol{p}_i$ is the $i$-th document vector in the first space measuring query-document similarity, and $\hat{\boldsymbol{p}}_i$ represents the $i$-th document vector in the second space quantifying political similarity between candidates.

In our implementation, we utilize AnglE-generated embeddings for the first space to measure query-document relevance, while the second space employs method-specific embeddings to capture political diversity. The trade-off between diversity and relevancy is controlled by the hyperparameter $\mu \in (0, 1)$, with $\lambda$ fixed at 0.5 across all experiments. The implementation of our modified D$k$MIPS algorithm is available at `https://github.com/dukesun99/PyDkMIPS`.

[3] `https://scikit-learn.org/1.5/modules/generated/sklearn.svm.SVC.html`

| Description | Symbol | Setting |
|---|---|---|
| Number of clusters | $k$ | 3,000 |
| Bias dispersion threshold (NewsSpectrum) | $\tau$ | 1.0 |
| Bias dispersion threshold (BigNews) | $\tau$ | 0.5 |
| Weighting factor for important topic retrieval | $\lambda$ | 0.8 |
| Minimum cluster size | $p$ | 50 |
| Number of topics (NewsSpectrum) | $\|M\|$ | 1,810 |
| Number of topics (BigNews) | $\|M\|$ | 2,279 |
| Learning rate | $\alpha$ | $10^{-6}$ |
| Batch size | $b$ | 4 |

Table 2: Hyperparameters used in our experiments.

## C LLM Generation Quality

To further illustrate the quality of the generated topics, we include four examples in Table 3. These topic dimensions exhibit clear, ideologically coherent indicators aligned with widely recognized partisan perspectives.

## D Additional Experiments

To further evaluate the generalizability and robustness of PRISM, we conduct three additional experiments: (1) classification on a human-annotated dataset (BASIL), (2) an alignment of label schemes for consistent cross-dataset comparison, and (3) a comparison against zero-shot LLM baselines.

**Classification on BASIL.** We evaluate PRISM on the BASIL dataset (Fan et al., 2019), a human-annotated corpus that labels each article as *Liberal*, *Center*, or *Conservative*. Notably, BASIL is entirely disjoint from any data used to train PRISM

| Generated Topic | Left Indicator | Right Indicator |
|---|---|---|
| U.S. Funding and Relationship with the WHO | Criticism of U.S. withdrawal, emphasis on global cooperation; concerns that defunding undermines pandemic response; calls for accountability from WHO without cutting ties. | WHO accused of being "China-centric," calls for defunding until reforms are made; WHO's handling of the pandemic criticized as ineffective; demands for the resignation of WHO leadership. |
| Disappearance and Violence in Mexico | Emphasis on femicide, human rights violations, and environmental activism; criticizes government inaction and suggests increasing violence against marginalized groups. | Focus on cartel violence and law enforcement failures; assertion of a need for stronger government action against organized crime and support for pro-family values amidst crisis. |
| Driver's License Regulations and Data Privacy | Emphasis on civil rights, privacy concerns over data selling by DMVs, criticism of government surveillance, and policies that support undocumented immigrants. | Focus on national security, criticism of relaxed immigration and driving laws for undocumented individuals, and concerns about the criminal misuse of counterfeit IDs. |
| Electric Vehicles and Alternative Powertrains | Advocates for government support, stricter regulations on emissions, and innovation in sustainable technology. | Emphasizes market-driven solutions, skepticism about government overreach and regulations, and preference for traditional vehicles. |

Table 3: Examples of generated topics and bias indicators.

| Model | Acc ↑ | Pre ↑ | Rec ↑ | F1-Ma ↑ | F1-Mi ↑ |
|---|---|---|---|---|---|
| **AnglE** | 35.0 | 28.8 | 28.9 | 28.8 | 34.7 |
| **Instructor** | 31.7 | 25.0 | 25.6 | 25.1 | 30.4 |
| **InBedder** | 33.3 | 29.0 | 26.7 | 26.8 | 32.8 |
| **CQG-MBQA** | 30.0 | 19.2 | 21.1 | 20.0 | 28.3 |
| **POLITICS** | 31.7 | 28.9 | 28.9 | 28.9 | 31.5 |
| **PRISM** | **40.0** | **38.5** | **36.7** | **37.3** | **39.7** |

Table 4: Classification results on BASIL.

| Model | Acc ↑ | Pre ↑ | Rec ↑ | F1-Ma ↑ | F1-Mi ↑ |
|---|---|---|---|---|---|
| **AnglE** | 62.7 | 61.6 | 58.3 | 58.9 | 62.7 |
| **Instructor** | 60.4 | 59.5 | 55.4 | 55.9 | 60.4 |
| **InBedder** | 64.1 | 66.7 | 58.3 | 59.0 | 64.1 |
| **CQG-MBQA** | 56.0 | 55.7 | 51.6 | 52.2 | 56.0 |
| **POLITICS** | 66.3 | 67.0 | 60.7 | 61.4 | 66.3 |
| **PRISM** | **92.8** | **91.2** | **92.8** | **91.8** | **92.8** |

Table 5: 3-class classification results on NewsSpectrum.

or its baselines, offering a strong test of out-of-distribution generalization.

For this experiment, we train PRISM on a mixture of the BigNews and NewsSpectrum datasets, then apply a logistic regression classifier to predict the political stance of BASIL articles using PRISM embeddings. As shown in Table 4, PRISM achieves the highest classification performance among all compared models, including POLITICS. These results further highlight the generalizability and robustness of our PRISM framework.

**Aligning Classification Labeling Scheme.** To facilitate consistent evaluation across datasets, we align the label schemes used in NewsSpectrum and BigNews. While NewsSpectrum employs a five-point ideological scale (*Left*, *Lean Left*, *Center*, *Lean Right*, *Right*), BigNews uses a simpler three-class schema (*Left*, *Center*, *Right*). We consolidate the five-point labels by mapping both *Left* and *Lean Left* to *Left*, and similarly merging *Right* and *Lean Right* into *Right*. Table 5 reports the performance under this aligned three-label setting, where PRISM continues to show strong results, confirming its stability under varying label granularities.

| Model | Acc ↑ | Pre ↑ | Rec ↑ | F1-Ma ↑ | F1-Mi ↑ |
|---|---|---|---|---|---|
| **Llama-3.1-8B** | 27.6 | 37.8 | 26.9 | 21.6 | 27.6 |
| **GPT-4o-mini** | 32.7 | 45.4 | 32.5 | 30.1 | 32.7 |
| **GPT-4o** | 39.1 | 49.8 | 38.7 | 38.7 | 39.1 |
| **PRISM** | **86.1** | **86.5** | **86.2** | **86.2** | **86.1** |

Table 6: Classification results on NewsSpectrum with three LLM zero-shot learning baselines.

**LLM Zero-shot Learning Baselines.** To contextualize PRISM's performance, we evaluate zero-shot political bias classification using powerful LLMs, including Llama-3.1-8B, GPT-4o-mini, and GPT-4o, on the NewsSpectrum dataset.

As shown in Table 6, while LLMs exhibit moderate performance out of the box, PRISM significantly outperforms them, particularly in Classification Accuracy (Acc) and F1-Macro (F1-Ma). These findings underscore the value of PRISM's domain-specific architecture: although LLMs are versatile, PRISM yields more accurate and interpretable results through its targeted representation of political bias.