When Bias Detection Breaks: The limits of supervised models across news sources

Juan Ignacio Llaberia

2026

Abstract

Political bias detection in news articles is often evaluated using random train–test splits, which can allow models to exploit publisher-specific artifacts rather than learn genuine ideological signals. We study the impact of this issue by comparing random splits with an outlet-controlled split where publishers are disjoint across training and test sets. Using 37,554 U.S. news articles labeled by outlet bias, we evaluate traditional machine learning models with sentence embeddings, a fine-tuned Transformer (ModernBERT), and zero- and few-shot prompting with a large language model. Results show that supervised models degrade sharply under outlet-controlled evaluation, with traditional approaches near chance and Transformers achieving only modest gains. In contrast, large language models generalize better to unseen outlets. These findings demonstrate that random splits overestimate performance and highlight the importance of outlet-controlled evaluation for robust political bias detection.

1 Introduction

Detecting political bias in news articles is an important task for media analysis, misinformation research, and information transparency. Automatic bias classification systems can help readers better understand ideological framing and enable large-scale studies of media behavior. Recent advances in machine learning, particularly Transformers and large language models, have significantly improved performance on many text classification tasks, making them attractive approaches for bias detection.

However, common evaluation practices may overestimate true model capability. Many studies rely on random train–test splits, where articles from the same news outlet appear in both training and evaluation sets. Because political bias labels are often derived from outlet-level ratings, models can learn publisher-specific stylistic patterns rather than genuine linguistic or ideological cues. This shortcut learning leads to inflated accuracy that may not generalize to unseen sources.

In this work, we systematically analyze the impact of data splitting strategies on political bias detection. Using a dataset of 37,554 U.S. news articles labeled by outlet bias, we compare two settings: a random split and a stricter outlet-controlled split that enforces publisher separation across partitions. We evaluate three modeling paradigms: traditional machine learning with sentence embeddings, a fine-tuned Transformer, and zero- and few-shot prompting with a large language model.

Our results show that performance drops substantially when models are evaluated on unseen outlets, confirming the presence of shortcut learning under random splits. While supervised approaches struggle to generalize, large language models demonstrate stronger robustness, suggesting that broad pretraining better captures ideological framing. These findings highlight the importance of careful dataset design and realistic evaluation protocols for reliable political bias detection.

2 Related Work

Political bias detection. Prior work has approached political bias classification using both traditional machine learning models with lexical features and more recent neural architectures, including contextual embeddings and Transformer-based models. These methods demonstrate that linguistic and stylistic cues can be used to infer ideological orientation, but most evaluations assume randomly split datasets, which may overestimate generalization performance.

Spurious correlations and shortcut learning. Recent studies in text classification show that models frequently exploit dataset artifacts, such as author or source-specific patterns, rather than task-relevant signals. This shortcut learning can lead to inflated results when train and test sets share superficial characteristics. Controlling for such leakage is therefore essential when evaluating real-world robustness.

Large language models for classification. Instruction-tuned large language models have recently shown strong zero- and few-shot performance across many NLP tasks, including document classification, without task-specific fine-tuning. Their broad pretraining and contextual reasoning capabilities suggest improved generalization under distribution shift.

Our contribution. Building on these findings, we explicitly study generalization in political bias detection by comparing outlet-controlled and random splits to quantify shortcut learning effects. We further provide a unified comparison between traditional machine learning, fine-tuned Transformers, and zero- and few-shot LLM prompting under the same evaluation protocol.

3 Hypotheses

H1. Models evaluated on outlet-controlled splits will show substantially lower performance than on random splits, indicating that random splits allow models to exploit publisher-specific artifacts rather than learning article-level ideological signals.

H2. Zero-shot prompted large language models will generalize better to unseen outlets than supervised approaches, including traditional machine learning and fine-tuned transformers, due to their broad pretraining and stronger contextual understanding of ideological framing.

4 Data

4.1 Dataset Overview

The dataset [1] consists of 37,554 news articles from 479 U.S. media outlets, labeled according to political bias as left, center, or right. The class distribution is relatively balanced, with the center class containing approximately 2,000 fewer samples.

Bias labels are derived from the political orientation of the publishing outlet, as defined by AllSides, rather than from article-level annotations. As a result, allowing articles from the same outlet to appear across multiple splits may enable shortcut learning, where models rely on publisher-specific style or metadata instead of learning linguistic indicators of bias. To address this issue, we construct two dataset variants: an outlet-based split (Section 4.2) and a random-based split (Section 4.3).

Each article includes metadata such as source name, URL, and topic. These attributes were excluded from model inputs to reduce the risk of learning spurious correlations. Although some topics exhibit partial association with political bias, our embedding-based representations

(see Section 4) are expected to implicitly encode topical information, making explicit inclusion unnecessary and potentially harmful.

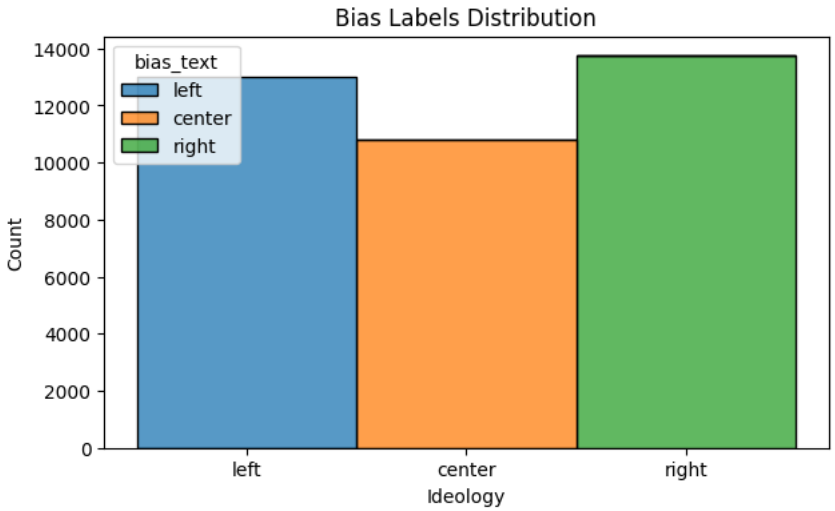

Figure 1: Distribution of samples across dataset splits. The dataset contains 29,258 training instances, 6,996 validation instances, and 1,300 test instances.

4.2 Outlet-Based Dataset

In the outlet-based configuration, all articles from a given news outlet are assigned exclusively to a single split (train, validation, or test).

Purpose: To prevent shortcut learning based on outlet-specific writing style or metadata.

Significance: This setting provides a stricter and more realistic evaluation of a model's ability to generalize to unseen sources.

4.3 Random-Based Split

In the random-based configuration, articles are shuffled and assigned to splits without regard to their source.

Purpose: To estimate an upper bound on model performance.

Significance: While this approach typically yields higher accuracy, the results may be inflated due to memorization of outlet-specific patterns rather than generalized bias detection.

5 Experiments

5.1 Traditional Machine Learning (XGBoost and Logistic Regression)

In this experiment, we train, evaluate, and analyze two traditional machine learning models: XGBoost and Logistic Regression. Both models are evaluated on the outlet-based and random-based dataset variants.

We select these architectures as strong and interpretable baselines. XGBoost provides a competitive non-linear tree-based approach that often performs well on structured representations, while Logistic Regression serves as a linear model that is well aligned with the geometric properties of high-dimensional embedding spaces and frequently performs strongly with sentence embeddings.

All articles are first embedded using `all-mpnet-base-v2`, a widely adopted Sentence-BERT model that produces reliable semantic representations for English text. For each dataset, we follow the same procedure. We first train baseline models using default hyperparameters to establish reference performance. We then perform hyperparameter optimization to identify improved configurations. Final models are evaluated on the held-out test set.

5.2 Fine-Tuned Transformer (ModernBERT)

For this stage, we fine-tune ModernBERT [2], a recent evolution of the BERT architecture trained on approximately two trillion tokens. The model supports input sequences of up to 8,192 tokens, making it well suited for long-form documents such as news articles.

We set the maximum sequence length to 2,048 tokens, which provides a practical trade-off between contextual coverage and computational cost. This limit captures the full content of most articles while maintaining efficient training. We expect contextual Transformer representations to outperform the traditional machine learning baselines.

Experimental Design

- **Fine-tuning strategy:** Rather than relying on static embeddings, we perform partial fine-tuning. Most Transformer layers are frozen to preserve general linguistic knowledge, while only the top two layers and the classification head are updated.
- **Hyperparameter optimization:** We conduct a search over key training parameters, particularly learning rate and weight decay, to ensure stable convergence.
- **Evaluation:** The resulting models are evaluated on the same test sets and directly compared with the XGBoost and Logistic Regression baselines.

5.3 Zero-Shot and Few-Shot LLM Prompting

In this setting, we evaluate a pre-trained large language model without additional training. Specifically, we use Qwen2.5-7B-Instruct [3], an instruction-tuned model designed for strong reasoning and text classification performance. We hypothesize that its prior knowledge and ability to analyze tone, semantics, and framing enable effective bias identification without task-specific fine-tuning.

Two prompting strategies are evaluated:

1. **Zero-shot:** The model receives only a task description and a single structured prompt, relying entirely on its pre-trained knowledge.
2. **Few-shot:** The same prompt is extended with one labeled example per class (left, center, right) to provide minimal task guidance.

Data contamination is an important consideration when evaluating LLMs. The model's knowledge cutoff predates the public release of this dataset, making leakage unlikely.

5.4 Evaluation Metrics

All experiments are evaluated using Accuracy, Precision, Recall, and F1-score. For LLM-based experiments, we additionally report the prediction error rate, defined as the proportion of outputs that fail to produce a valid label. This metric ensures that reported performance reflects only well-formed predictions.

6 Results

We evaluate all models under both splitting strategies to quantify the impact of outlet leakage on performance and generalization. Results are reported using Accuracy, Precision, Recall, and Macro F1-score. Confusion matrices are included to analyze class-level behavior and error patterns.

6.1 Outlet-Based Dataset

Traditional Machine Learning (XGBoost and Logistic Regression) Both XGBoost and Logistic Regression achieve near-chance performance when evaluated on previously unseen outlets. Macro F1-scores remain close to 0.33, indicating that the models fail to generalize beyond the publishers observed during training. The strong validation performance obtained during hyperparameter optimization does not transfer to the test set, suggesting reliance on outlet-specific artifacts rather than article-level ideological signals.

Metric	Value
Accuracy	0.3422
Precision	0.3528
Recall	0.4005
Macro F1-Score	0.3386
Log Loss	2.4223
Inference Time	0.06s

Table 1: XGBoost/Logistic Regression Performance (Outlet-Based)

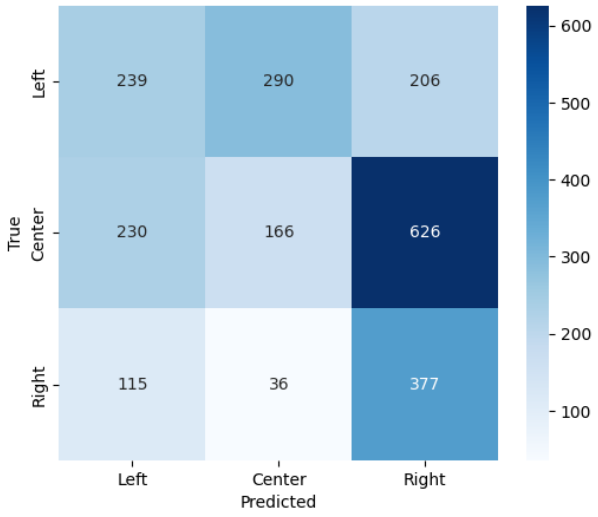

Figure 2: Confusion matrix for XGBoost/Logistic Regression on the outlet-based split. The confusion matrix shows substantial mixing across all classes.

Fine-Tuned Transformer (ModernBERT) ModernBERT improves upon the traditional baselines but still struggles under strict outlet separation. While contextual representations provide moderate gains (Macro F1 $\approx 41\%$), performance remains limited, indicating that ideological bias detection remains challenging without access to source-specific cues.

Metric	Value
Accuracy	43.37%
Precision	45.28%
Recall	43.37%
F1-Score	41.33%
Loss	1.2921
Throughput	21.96 samples/s

Table 2: ModernBERT Performance (Outlet-Based)

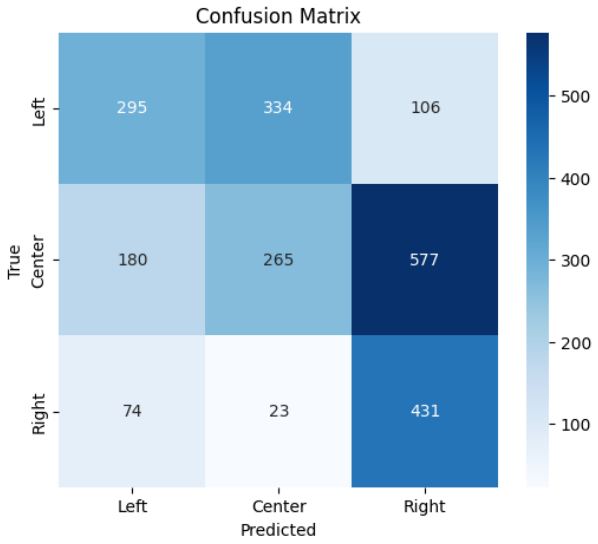

Figure 3: Confusion matrix for ModernBERT on the outlet-based split. Substantial cross-class confusion persists.

6.2 Random-Based Dataset

Traditional Machine Learning (XGBoost and Logistic Regression) In contrast, the same models achieve substantially higher performance under the random split. Because articles from the same outlets appear in both training and test sets, the models can exploit consistent stylistic or publisher-specific patterns, resulting in large improvements across all metrics.

Metric	Value
Accuracy	0.5900
Precision	0.5668
Recall	0.5663
Macro F1-Score	0.5665
Log Loss	0.8866
Inference Time	0.05s

Table 3: XGBoost/Logistic Regression Performance (Random-Based)

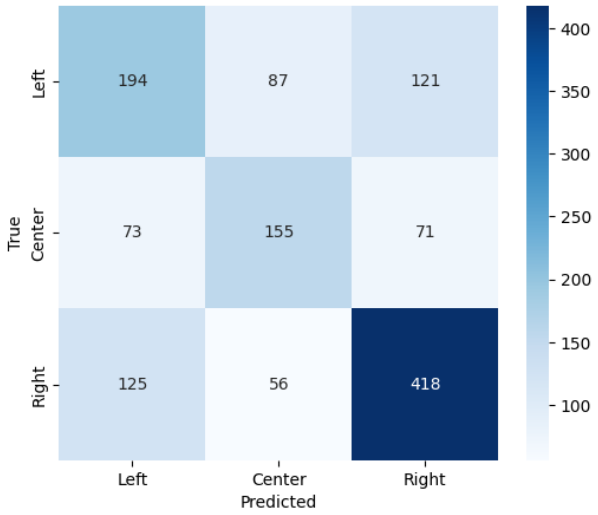

Figure 4: Confusion matrix for XGBoost/Logistic Regression on the random-based split.

Fine-Tuned Transformer (ModernBERT) Performance increases dramatically for the Transformer under the random split, exceeding 90% across all metrics. Although these results appear highly competitive, they likely reflect memorization of outlet-specific characteristics rather than genuine ideological reasoning.

Metric	**Value**
Accuracy	90.08%
Precision	90.36%
Recall	90.08%
F1-Score	90.06%
Loss	0.2570
Throughput	22.12 samples/s

Table 4: ModernBERT Performance (Random-Based)

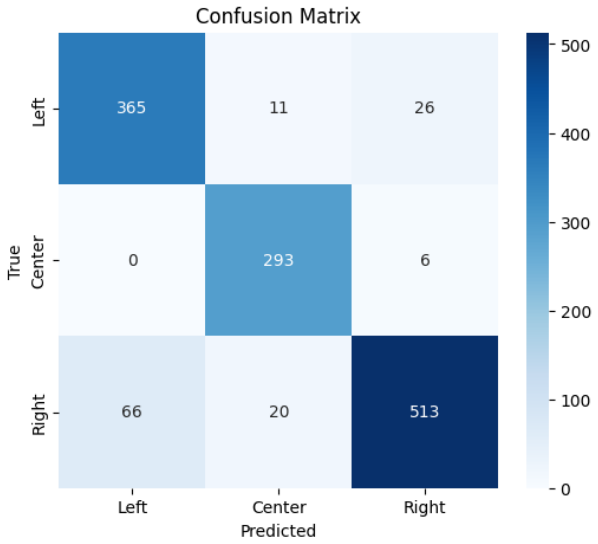

Figure 5: Confusion matrix for ModernBERT on the random-based split. The near-perfect diagonal indicates that the model can reliably distinguish outlets present in both splits.

6.3 Large Language Model Results

Zero-Shot Prompting The zero-shot setting achieves approximately 60% accuracy, substantially outperforming both traditional machine learning models (∼33%) and the fine-tuned Transformer (∼45%) on the outlet-based split. This suggests that broad pretraining and instruction tuning enable the model to capture higher-level semantic and contextual cues related to ideological framing without task-specific supervision.

Metric	Value
Accuracy	59.80%
Precision	58.54%
Recall	55.77%
F1-Score	53.21%
Error Rate	2.31%
Avg. Latency	3.30s / article

Table 5: Zero-Shot Prompting Results

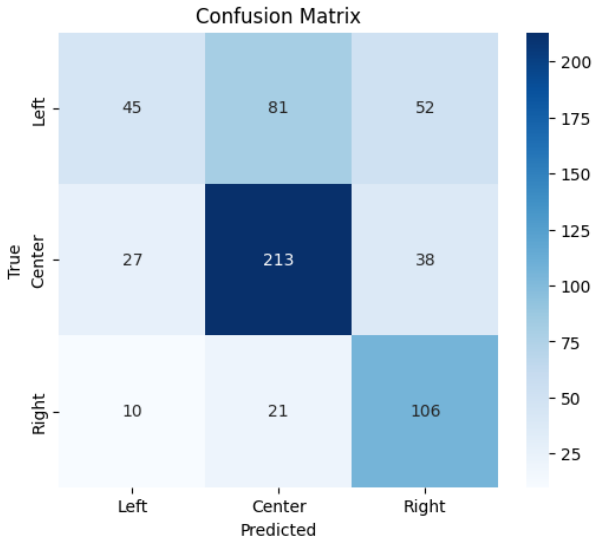

Figure 6: Confusion matrix for zero-shot prompting.

Few-Shot Prompting (3-shot) Few-shot prompting further improves performance, surpassing 70% accuracy and achieving the highest scores among all outlet-controlled experiments. However, this improvement comes at the cost of increased inference time, with latency approximately three times higher than zero-shot prompting.

Metric	Value
Accuracy	70.79%
Precision	73.32%
Recall	68.20%
F1-Score	69.93%
Error Rate	0.00%
Avg. Latency	9.39s / article

Table 6: Few-Shot Prompting Results

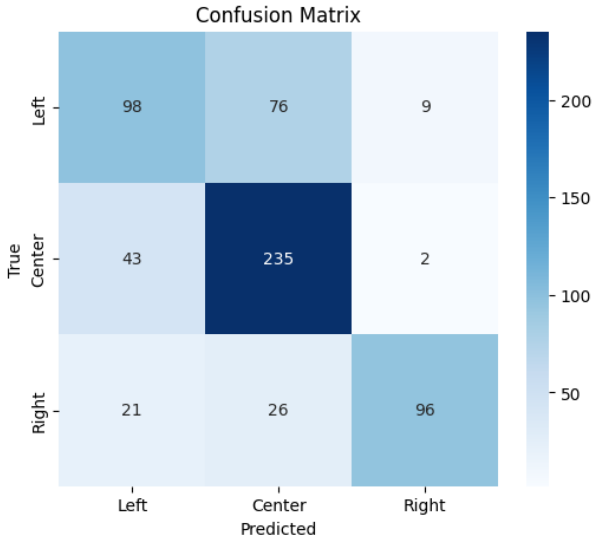

Figure 7: Confusion matrix for few-shot prompting.

6.4 Summary of Findings

Across all architectures, performance drops sharply when evaluated on unseen outlets. This consistent degradation indicates that random splits introduce shortcut learning and substantially overestimate real-world performance. While fine-tuned Transformers improve over traditional baselines, zero- and few-shot LLM prompting demonstrates the strongest generalization under strict outlet separation. These findings highlight the importance of outlet-controlled evaluation when measuring robust political bias detection.

7 Conclusion

This work systematically evaluated political bias detection under two data splitting strategies to measure the impact of outlet leakage on model generalization. Across all architectures, performance decreased substantially when evaluated on unseen outlets compared to random splits. This consistent degradation indicates that models trained with random splits exploit publisher-specific stylistic patterns, leading to shortcut learning and inflated performance estimates that do not reflect true ideological understanding.

Under the stricter outlet-controlled setting, traditional machine learning models performed near chance, while fine-tuned Transformers achieved only moderate improvements. In contrast, zero-shot and few-shot prompting with large language models demonstrated stronger generalization, suggesting that broad pretraining enables these models to capture higher-level semantic and contextual cues related to framing and tone rather than relying solely on source-specific artifacts.

Despite these advantages, LLM-based approaches introduce practical limitations, including higher computational cost, increased latency, and potential risks of implicit data leakage due to large-scale pretraining. Therefore, while LLMs show promise for robust bias detection, their deployment requires careful evaluation and resource considerations.

Future work will focus on scaling the study to larger and more diverse datasets, extending the analysis to additional countries and media ecosystems to assess cross-cultural generalization, and exploring fine-tuning strategies for large language models to combine their reasoning capabilities with task-specific supervision. We also plan to investigate methods that better separate ideological signals from publisher artifacts to further reduce shortcut learning and improve robustness.

References

[1] Siddharth M. B., Article Bias Prediction Dataset. Hugging Face Datasets, 2025. https://huggingface.co/datasets/siddharthmb/article-bias-prediction-all

[2] Hugging Face, ModernBERT Documentation. Transformers Model Docs, 2024. https://huggingface.co/docs/transformers/model_doc/modernbert

[3] Qwen Team, Qwen2.5-7B-Instruct. Hugging Face Models, 2024. https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[4] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of EMNLP-IJCNLP*.

[5] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[6] Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society*.

[7] AllSides. Media Bias Ratings Methodology. https://www.allsides.com/media-bias/media-bias-ratings

[8] Brown, T. et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems (NeurIPS)*.

[9] Llaberia, J. (2026). Political Bias Detection Experiments (Code Repository). GitHub. https://github.com/JuaniLlaberia/news-bias-detection-research