

Detecting Framing Bias in News via Probabilistic Graphical Modeling

1st Shirin Shujaa
SSET, RMIT University
HCMC, Vietnam
0009-0007-7408-3848

2nd Ginel Dorleon
SSET, RMIT University
HCMC, Vietnam
0000-0003-2343-4445

Abstract—Framing bias in news is a subtle yet impactful form of media bias where journalists shape a narrative through selective presentation and wording. Detecting such bias automatically is challenging, as it requires understanding beyond overt partisan cues. In this paper, we propose a novel approach for detecting framing bias in news text using a probabilistic graphical model. In contrast to recent chain-of-thought prompting with large language models, our method leverages a Conditional Random Field (CRF) to model subtle linguistic and structural cues of framing bias across article sentences. We formalize the framing bias detection task on the Media Bias Identification Benchmark (MBIB) and describe our model’s mathematical underpinnings. Experiments on MBIB show that our approach outperforms fine-tuned BERT and ConvBERT baselines in F1-score indicating superior ability to recognize nuanced framing tactics. We present ablation studies demonstrating the contribution of lexical bias features and cross-sentence dependencies, and provide qualitative analysis of model decisions. Our findings highlight the value of structured probabilistic modeling for media bias detection. This work offers a new avenue for framing bias identification, paving the way for more transparent news consumption and robust, multi-faceted bias detection systems.

Index Terms—Framing bias, media bias, probabilistic graphical model, LLMs

I. INTRODUCTION

Media bias in news reporting takes many forms, with framing bias among the most subtle [1], [2]. Framing bias arises when facts are presented in ways that guide interpretation through emphasis or emotive wording without altering the facts themselves. Unlike ideological bias such as left vs. right, framing works at the level of narrative and tone, often through word choice or connotation (for example: “government’s heartless cutbacks” vs. “funding reduction”). Such language can sway opinion and erode media trust [3], [4]. Automatic detection is therefore crucial for both readers and balanced aggregation.

Most computational work has addressed coarse bias like ideology or source-level polarity. For example, SemEval-2019 on hyperpartisan news targeted left/right bias [5], with many systems treating bias prediction as a text classification problem [6], [7]. Transformer-based models like BERT achieved strong results for partisan bias [6], [8]. Framing bias, however, is less studied: it involves subtle semantic cues or omissions beyond partisan keywords [9], [10]. Lexicons and topic models detect some biased terms but fail on context. LLMs attempt

zero/few-shot detection, yet often confuse emotional tone with bias; even GPT-4 with chain-of-thought struggles without explanations [11].

We address this gap with a probabilistic graphical model. Specifically, we introduce a Conditional Random Field (CRF) that models sentences as a sequence with bias labels, capturing lexical cues and cross-sentence dependencies [2], [12]. For instance, an article that uses sympathetic language for one actor and critical language for another can be captured through linked sentence-level indicators. We evaluate on the Media Bias Identification Benchmark (MBIB) [13], focusing on linguistic bias tasks that involve framing by word choice and tone [7], [13].

Our contributions are as follows.

- 1) We formulate framing bias detection as a structured prediction task and introduce a CRF-based model to address it.
- 2) We design lexical and semantic features (sentiment, subjectivity, bias lexicon counts) and integrate them into a graphical model for the first time, in contrast to prior end-to-end neural methods.
- 3) Through comprehensive experiments on the MBIB dataset, we show that our model outperforms strong baselines including fine-tuned BERT, ConvBERT, and a GPT-3.5 prompt-based classifier. Our approach is competitive, especially in precision on subtle bias cases.
- 4) We provide ablation studies and qualitative analysis showing how the model captures cross-sentence bias consistency and rare bias cues. To our knowledge, this is the first probabilistic graphical model applied to framing bias detection.

The remainder of this paper is organized as follows. Section II reviews related work on media bias detection, highlighting methods for bias by word choice and framing. Section III details our proposed CRF-based model and its theoretical formulation. Section IV describes the experimental setup, dataset preparation from MBIB, baseline comparisons and results. Section V concludes with future directions.

II. RELATED WORK

Media bias has been studied widely in NLP and computational social science [2], [3]. Early work examined lexical bias, loaded language signaling prejudice or favoritism. Recasens

et al. [10] used linguistic features to detect biased Wikipedia edits, and Fan *et al.* [9] introduced the BASIL dataset with sentence-level annotations, showing framing bias is often subtler than opinion. Lexicons and handcrafted cues [9], [10] can flag biased words but fail on context or discourse.

Neural models shifted the field toward document classification for ideology prediction [6], [7]. Baly *et al.* [6] fine-tuned BERT to predict outlet ideology; Spinde *et al.* [7] leveraged the BABE dataset with distant supervision. SemEval-2019 on hyperpartisan news [5] further cemented focus on ideological polarity, where transformers excelled. Yet, strong performance on political bias does not extend to framing, which relies on linguistic nuance [3], [4].

For framing and subtle bias, specific efforts on framing include Corman and Liu [14], who used handcrafted features across outlets. More recently, Pastorino *et al.* [11] applied prompting with GPT-3.5/4 to detect framing in headlines; chain-of-thought improved consistency but models still conflated emotional tone with bias. These highlight the limits of LLM prompting.

Alternatives to end-to-end transformers include probabilistic and structured approaches. Chen *et al.* [15] modeled sentence-level variation with Gaussian bias distributions. Hofmann *et al.* [12] used graph neural networks to capture ideological framing across discourse. Krieger *et al.* [8] achieved state-of-the-art with domain-adaptive RoBERTa. Others tackled mitigation: Lee *et al.* [16] generated “neutralized” versions of articles, while Pryzant *et al.* [17] proposed automatic debiasing frameworks.

The MBIB benchmark [13] unified bias detection tasks, including linguistic bias central to framing. Results showed ConvBERT and RoBERTa perform well on political bias but lag on linguistic bias [13], motivating structured, feature-driven approaches like ours over black-box classifiers.

III. METHODOLOGY

In this section, we formalize the framing bias detection task and describe our Conditional Random Field (CRF) model. We first define the problem setting, then detail the probabilistic model and features, and finally outline the learning and inference procedures.

A. Problem Formulation

We represent each news article as a sequence of sentences $\mathbf{x} = (x_1, x_2, \dots, x_N)$. The task is to detect framing bias not with a single article-level label, but by assigning each sentence x_i a binary bias label $y_i \in \{0, 1\}$, where $y_i = 1$ denotes biased framing and $y_i = 0$ denotes neutrality. An article is considered biased if any sentence is biased. This formulation follows prior work showing that biased articles often mix neutral and biased statements, and that identifying these “bias spans” is informative for overall bias [2], [9], [10].

Formally, let $\mathbf{y} = (y_1, \dots, y_N)$ be the latent bias label sequence. We model its conditional probability given text \mathbf{x} with a linear-chain CRF:

$$P_{\theta}(\mathbf{y} | \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \left(\sum_{i=1}^N \sum_k \theta_k f_k(y_{i-1}, y_i, \mathbf{x}, i) \right) \quad (1)$$

where $Z(\mathbf{x})$ is the normalization term, $\{f_k\}$ are feature functions, and θ_k their weights [15]. Features can depend on the current sentence x_i , its label y_i , and the previous label y_{i-1} , enabling both state and transition features. Intuitively, the CRF integrates (1) how likely each sentence is biased given its text, and (2) how bias states transition across sentences. The latter captures patterns such as biased sentences clustering together, or contextual continuity making nearby sentences more likely to share bias.

B. Feature Design

We design the feature functions f_k to capture linguistic cues of framing bias:

- **Lexical Polarization Features:** Using a sentiment lexicon and the MPQA subjective clues lexicon¹, we count polarized words in x_i . For instance, “heroic” or “disastrous” activate $f_{\text{pos}}(y_i, \mathbf{x}, i)$ or $f_{\text{neg}}(y_i, \mathbf{x}, i)$. Word choice such as “illegal alien” vs. “undocumented immigrant” reflects framing [9], [10]. We expect $y_i = 1$ (biased) to correlate with higher counts.

- **Bias Lexicon Features:** We compiled bias-indicating terms from prior work [9], [10] and journalism guidelines (for example, “terrorist” vs. “freedom fighter”, honorifics like “so-called”). A feature f_{biasLex} counts such terms in x_i , capturing classic framing choices.

- **Modal and Attributive Features:** Framing often relies on hedging, scare quotes, or weasel words [3]. We include features for modal verbs and phrases (“might indicate”, “reportedly”), as well as quotation usage, signaling uncertainty or emphasis.

- **Syntactic/Position Features:** We add a feature for sentence position, since framing often appears in the lead or conclusion [2], [14]. We also check if a sentence is quoted speech. For example, $f_{\text{lead}}(y_i)$ activates if $i = 1$.

- **Contextual Embedding Features:** Alongside symbolic cues, we use BERT sentence embeddings (classification token). Features are defined as components of a reduced embedding vector when y_i takes a value, allowing the CRF to capture latent semantics. For example, a sentence linking immigration to crime may produce an embedding indicative of framing.

The full feature vector for sentence i is:

$$f(y_{i-1}, y_i, \mathbf{x}, i) = [\mathcal{K}\{y_{i-1}, y_i\}, \text{LexCount}(x_i), \text{BiasLex}(x_i), \text{Modal}(x_i), \text{Position}(i), \text{Embed}(x_i)] \quad (2)$$

where $\mathcal{K}\{y_{i-1}, y_i\}$ indicates label transitions and $\text{Embed}(x_i)$ is the reduced embedding. Model parameters θ include weights for transitions (capturing clustering of bias, i.e. $y_{i-1} = 1, y_i = 1$) and for lexical/embedding features under $y_i = 0$ or $y_i = 1$.

¹https://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

For instance, the CRF can learn high weights for sentiment words under bias and near-zero under neutral.

C. Model Training and Inference

We train the CRF using supervised learning on MBIB, which provides binary sentence-level bias labels across datasets such as BASIL [9], Wikipedia Neutrality Corpus, and Media Frames Corpus [13]. A sentence is labeled 1 if it contains linguistic bias (framing, loaded language, stereotypes) and 0 otherwise [3]. MBIB binarizes heterogeneous definitions for consistency. Training optimizes the conditional log-likelihood:

$$\mathcal{L}(\theta) = \sum_j \log P_\theta(\mathbf{y}^{(j)} | \mathbf{x}^{(j)}) - \lambda \|\theta\|^2 \quad (3)$$

with L_2 regularization to avoid overfitting given the large feature set. The gradient is computed via forward-backward [15], and parameters θ are learned with stochastic gradient descent. To accelerate convergence, we initialize embedding feature weights from a BERT-based bias classifier trained separately.

For inference on a new article, we predict $\hat{\mathbf{y}} = \arg \max_{\mathbf{y}} P_\theta(\mathbf{y}|\mathbf{x})$ using the Viterbi algorithm [18], which runs in $O(N \cdot |\mathcal{Y}|^2)$ time with $|\mathcal{Y}| = 2$, i.e. linear in article length N [15]. Each sentence is tagged as biased or neutral, and an article is classified as biased if at least one $\hat{y}_i = 1$.

The advantage here is interpretability. Feature weights θ reveal which cues are linked to bias. For example, a strongly positive weight for “lead sentence with $y = 1$ ” means the model learned that biased framing often appears in the opening. Likewise, word-level and embedding dimensions can be inspected as bias signals, making results explainable.

Computation is modest: feature extraction involves lexicon lookups and a BERT embedding per sentence, while inference is linear in N . In practice, inference took milliseconds per article, dominated by embedding computation. Training on a few thousand articles converged within a few epochs. This makes the method suitable for large-scale bias analysis, complementing recent work on scaling detection across outlets [19].

IV. EXPERIMENTS

We evaluate our approach on the Media Bias Identification Benchmark (MBIB) [13], focusing on framing and linguistic bias. We compare against baselines and report both quantitative and qualitative results.

A. Dataset and Setup

MBIB [13] unifies 22 datasets into 9 bias detection tasks. We focus on the linguistic bias task, which captures word- and sentence-level phenomena such as framing and connotation bias. Representative datasets include BASIL [9], the Wikipedia Neutrality Corpus [10], Media Frames Corpus (MFC) [20], Biasly Annotated News [21], and RedditBias [22].

We follow MBIB preprocessing protocols and adopt the official train/validation/test splits. Evaluation metrics are Accuracy, F1-score for the biased class, Precision, and Recall.

B. Main Comparison

Table I compares our CRF with strong baselines. The CRF achieves the best overall performance, with 88.6% accuracy and an F1 of 78.8, surpassing BERT, ConvBERT, and DA-RoBERTa. Gains are most pronounced in precision (80.1), indicating fewer false positives, while recall remains competitive. Zero-shot GPT-3.5 performs markedly worse, aligning with prior reports that LLMs underperform on subtle framing bias [11]. These results show that structured modeling adds value even against large neural baselines.

TABLE I
BIAS DETECTION PERFORMANCE ON MBIB (LINGUISTIC/FRAMING BIAS TASK). BEST RESULTS IN BOLD.

Model	Acc.	F1	Prec.	Rec.
Majority (Neutral)	81.2	0.0	–	–
BERT (fine-tuned)	86.5	75.4	77.1	73.8
ConvBERT (fine-tuned)	87.0	76.5	78.3	74.8
DA-RoBERTa (Krieger et al.)	87.3	77.0	79.5	75.2
GPT-3.5 (zero-shot)	80.4	69.8	65.2	75.0
Proposed CRF	88.6	78.8	80.1	77.6

C. Ablation Study

We ablated main features (Table II). Transition features contribute +1.1 F1, confirming the benefit of sequence modeling. Lexicon features add interpretability and improve F1 by 1.3, while embeddings provide the largest boost overall. Even without embeddings, the CRF remains competitive with ConvBERT, highlighting the complementary role of domain-specific features alongside contextual embeddings.

TABLE II
ABLATION OF FEATURES IN OUR CRF MODEL.

Model Variant	F1	Acc.
Full CRF model	78.8	88.6
- no Transition features	77.7	87.9
- no Lexicon features	77.5	88.0
- no Embedding features	77.3	87.5
- no Lexicon & no Embedding	75.6	86.8

D. Cross-Domain Generalization

To test robustness, we trained on BASIL [9] and tested on Wikipedia Neutrality [10], and vice versa (Table III). Neural baselines showed clear overfitting, with cross-domain F1 in the mid-60s. Our CRF consistently achieved above 70, a relative gain of 4-5 points. This indicates stronger generalization, likely because the model exploits lexical and transition patterns that transfer across domains rather than memorizing topic correlations.

E. Bias-Type Breakdown

Performance by bias subtype is shown in Table IV. The CRF excels on “Loaded Words” (78.0 F1) and “Emotive Language” (76.2), where lexicon and sentiment cues are critical. On “Framing by Omission,” all models perform poorly (≤ 63 F1),

TABLE III
CROSS-DOMAIN GENERALIZATION: TRAIN ON ONE DATASET, TEST ON ANOTHER.

Train → Test	BERT F1	CRF F1
BASIL → Wikipedia	67.2	71.9
Wikipedia → BASIL	65.8	70.5

reflecting the difficulty of detecting information that is absent. This breakdown highlights where structured models bring the most benefit and where new methods are needed.

TABLE IV
PERFORMANCE BY BIAS SUBTYPE WITHIN MBIB (LINGUISTIC BIAS TASK).

Subtype	BERT F1	ConvBERT F1	CRF F1
Loaded Words	72.4	74.1	78.0
Emotive Language	70.6	71.5	76.2
Framing by Omission	61.3	62.0	62.8
Overall	75.4	76.5	78.8

F. Error Analysis

We examined 50 articles misclassified by either BERT or our CRF and found three main patterns. In cases of subtle bias with only a few loaded phrases, the CRF was more sensitive; for instance, it flagged “reckless spending” as biased while BERT labeled the article neutral. BERT, however, often misclassified neutral texts on sensitive topics such as immigration, likely due to topic correlations in training data, whereas the CRF correctly identified them as neutral by relying on explicit features. The CRF struggled more on very short articles, sometimes over-labeling bias when consecutive negative words appeared, even if ground truth considered them appropriate in context. These findings suggest that incorporating metadata, such as article genre, could reduce such errors.

V. CONCLUSION

We introduced a CRF-based model for framing bias detection that leverages linguistic features and sequence structure across sentences. On MBIB’s linguistic bias tasks, our model surpassed strong baselines and provided interpretable sentence-level predictions.

These results show that probabilistic modeling is effective for capturing subtle framing bias often missed by neural classifiers. Future work will extend beyond binary detection to multi-class framing, such as distinguishing economic versus moral frames, through an expanded state space or hierarchical modeling. The full codebase is available [here](#).

REFERENCES

- [1] R. M. Entman, “Framing: Toward clarification of a fractured paradigm,” *Journal of Communication*, vol. 43, no. 4, pp. 51–58, 1993.
- [2] T. Spinde, S. Hinterreiter, F. Haak, T. Ruas, H. Giese, N. Meuschke, and B. Gipp, “The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias,” *ACM Computing Surveys*, 2023, in review.
- [3] F.-J. Rodrigo-Ginés, J. C. de Albornoz, and L. Plaza, “A systematic review on media bias detection: what is media bias, how it is expressed, and how to detect it,” *Expert Systems with Applications*, vol. 237, p. 121641, 2024.
- [4] P. Mavridis, O. Inel, X. Wilcke, M. Makhortykh, M. de Jong, H. Mazeppus, A. Dimitrova, J. de Vos, A. Bozzon, and T. Kuhn, “Framing is mightier than the sword: Detection of episodic and thematic framing in news media,” *Human Computation*, vol. 11, no. 1, pp. 1–28, 2024.
- [5] J. Kiesel, M. Mestre, R. Shukla, E. Vincent, D. Corney, P. Adineh, B. Stein, and M. Potthast, “Semeval-2019 task 4: Hyperpartisan news detection,” in *Proceedings of SemEval*, 2019, pp. 829–839.
- [6] R. Baly, G. D. S. Martino, J. Glass, and P. Nakov, “We can detect your bias: Predicting the political ideology of news articles,” in *Proceedings of EMNLP*, 2020, pp. 4982–4991.
- [7] T. Spinde, M. Plank, J.-D. Krieger, T. Ruas, B. Gipp, and A. Aizawa, “Neural media bias detection using distant supervision with BABE,” in *Findings of the Association for Computational Linguistics: EMNLP*, 2021, pp. 1166–1177.
- [8] J.-D. Krieger, T. Spinde, T. Ruas, J. Kulshrestha, and B. Gipp, “A domain-adaptive pre-training approach for language bias detection in news,” in *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2022.
- [9] L. Fan, M. White, E. Sharma, R. Su, P. K. Choubey, R. Huang, and L. Wang, “In plain sight: Media bias through the lens of factual reporting,” in *Proceedings of EMNLP-IJCNLP*, 2019, pp. 6343–6349.
- [10] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky, “Linguistic models for analyzing and detecting biased language,” in *Proceedings of the 51st Annual Meeting of ACL*, 2013, pp. 1650–1659.
- [11] V. Pastorino, J. A. Sivakumar, and N. S. Moosavi, “Decoding news narratives: A critical analysis of large language models in framing bias detection,” *arXiv preprint arXiv:2402.11621*, 2024.
- [12] V. Hofmann, X. Dong, J. B. Pierrehumbert, and H. Schütze, “Modeling ideological salience and framing in polarized online groups with graph neural networks,” in *Findings of the Association for Computational Linguistics: NAACL*, 2022, pp. 536–550.
- [13] M. Wessel, T. Horych, T. Ruas, A. Aizawa, B. Gipp, and T. Spinde, “Introducing MBIB – the first media bias identification benchmark task and dataset collection,” in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2023, pp. 21–30.
- [14] S. R. Corman and H. Liu, “Identifying framing bias in online news,” *ACM Transactions on Social Computing*, vol. 1, no. 2, pp. 5:1–5:18, 2018.
- [15] W.-F. Chen, K. Al-Khatib, B. Stein, and H. Wachsmuth, “Detecting media bias in news articles using gaussian bias distributions,” in *Proceedings of EMNLP*, 2020, pp. 8959–8970.
- [16] N. Lee, Y. Bang, A. Madotto, and P. Fung, “Mitigating media bias through neutral article generation,” *arXiv preprint arXiv:2104.00336*, 2021.
- [17] R. Pryzant, R. D. Martinez, N. Dass, S. Kurohashi, D. Jurafsky, and D. Yang, “Automatically neutralizing subjective bias in text,” *arXiv preprint arXiv:1911.09709*, 2019.
- [18] H.-L. Lou, “Implementing the viterbi algorithm,” *IEEE Signal processing magazine*, vol. 12, no. 5, pp. 42–52, 2002.
- [19] R. Rönneback, C. Emmery, and H. Brighton, “Automatic large-scale political bias detection of news outlets,” *PLOS ONE*, vol. 20, no. 5, p. e0321418, 2025.
- [20] H. Kwak, J. An, and Y.-Y. Ahn, “A systematic media frame analysis of 1.5 million New York Times articles,” in *Proceedings of the 12th ACM Conference on Web Science (WebSci)*, 2020, pp. 305–314.
- [21] P.-L. Huguët-Cabot, D. Abadi, A. Fischer, and E. Shutova, “Us vs. them: A dataset of populist attitudes, news bias and emotions,” in *Proceedings of EACL*, 2021, pp. 1921–1945.
- [22] S. Barikeri, A. Lauscher, and I. Vulić, “Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models,” in *Proceedings of AACL-IJCNLP*, 2021.