

Contrastive Sentiment Context for Sentence-Level Media Bias Detection

Group36

Abstract

Sentence-level media bias detection is challenging due to the subtle and implicit nature of bias in news writing. Prior work often relies on complex modeling architectures or dataset-specific annotations to capture contextual dependencies. In this work, we explore a simpler alternative by focusing on individual sentences and treating media bias as a relative evaluative phenomenon, reflected through sentiment contrast with surrounding context. We propose a contrastive sentiment context approach that selects contextual sentences with maximal evaluative difference from the target sentence. Our results show that sentiment-based contrastive context provides a consistent performance advantage across datasets and backbone encoders, indicating that emphasizing evaluative contrast helps the model better identify biased framing.

1 Introduction

Media bias in news reporting has long been recognized as a critical factor influencing public opinion and political attitudes. While biased reporting is often associated with explicit opinionated language, a substantial portion of media bias manifests implicitly through subtle framing choices, word selection, and emphasis. Recent work has therefore emphasized sentence-level media bias detection, as even individual sentences can contribute to biased interpretations of otherwise factual news articles (Fan et al., 2019; Lim et al., 2020).

From an NLP perspective, sentence-level media bias detection is commonly formulated as a classification task. However, such judgments are inherently contextual and relative: a sentence may appear neutral in isolation, yet convey bias when contrasted with surrounding sentences or with alternative descriptions of the same event. This relative nature of bias complicates the formulation of the task and raises questions about how contextual information should be incorporated.

Existing approaches address this challenge by explicitly modeling contextual structures, including neighboring sentence dependencies, discourse relations, event-level interactions, or by leveraging dataset-specific annotations to construct enriched contexts (Lei et al., 2022; Lei and Huang, 2024; Maab et al., 2023). While these methods demonstrate that context is essential for bias detection, they often rely on complex modeling pipelines or assume access to fine-grained annotations.

In this work, we operationalize contextual contrast using an evaluative sentiment signal, and examine whether relative differences in sentence-level sentiment within the same article can serve as a sufficient cue for biased framing.

Based on this formulation, we investigate the following research questions:

- **RQ1:** Can sentence-level media bias be captured using a sentence-centric approach that models bias as relative evaluative sentiment differences within an article?
- **RQ2:** How does sentiment-guided contrastive context modeling compare to sentence-only, naive window-based, and random context aggregation baselines across different sentence-level media bias datasets?
- **RQ3:** How does varying the number of sentiment-contrastive contextual sentences affect sentence-level media bias detection performance?

2 Related Work

Sentence-level Media Bias Datasets Early studies on media bias detection mainly focused on document-level or source-level analysis (Baly et al., 2018, 2020; Kiesel et al., 2019). More recent work has shifted attention to sentence-level media bias, motivated by the observation that biased framing at

the sentence level can significantly influence readers even when the overall article appears neutral. Representative datasets include BASIL (Fan et al., 2019), which provides fine-grained sentence-level annotations along with rich metadata such as bias type, target entities, and speaker information, and the dataset introduced by (Lim et al., 2020), which annotates sentence-level bias relative to reference articles without relying on detailed structural annotations. These datasets enable the study of subtle and implicit bias at a fine granularity, while also differing substantially in their annotation schemes and available supervision signals.

Contextual Modeling One line of research addresses sentence-level media bias detection by introducing complex model architectures that explicitly encode contextual structures beyond individual sentences. Such approaches model discourse relations or event-level interactions to capture long-range dependencies and implicit bias cues, demonstrating that structured contextual reasoning can improve bias detection performance (Lei et al., 2022; Lei and Huang, 2024). While effective, these methods typically require sophisticated model designs and additional supervision signals, increasing computational complexity and limiting their applicability across datasets.

Dataset-dependent Contextual Augmentation Another line of work incorporates contextual information by leveraging fine-grained annotations provided by specific datasets, particularly BASIL. These approaches exploit metadata such as bias type, target entities, or event information to construct bias-aware or target-aware contexts, enabling effective context augmentation for sentence-level bias detection (Maab et al., 2023). Although such strategies benefit from rich supervision and achieve strong results on BASIL, they heavily depend on the availability of detailed annotations, which may not generalize to datasets where only coarse sentence-level bias labels are available.

Sentiment-based Approaches Several studies have explored sentiment analysis as a signal for detecting media bias, motivated by the observation that biased reporting often departs from neutral tone through evaluative or emotionally charged language (Recasens et al., 2013). Early approaches commonly rely on sentiment polarity and subjectivity features, under the assumption that biased sentences tend to express stronger evaluative stance

Sentiment Contrast within a BASIL Article (HPO / Trump Debate)

Sentence 1 (Context, Unbiased)

Donald Trump has announced Tuesday that he is pulling out of his own Dec. 27th debate.

Sentiment: Neutral (score ≈ 0.02)

Sentence 2 (Context, Unbiased)

The announcement came just eleven days after Trump initially said he was partnering with the conservative magazine to host a forum with the remaining Republican presidential candidates.

Sentiment: Neutral (score ≈ 0.01)

Sentence 3 (Target, Biased)

Trump also attracted the usual share of public ridicule for his decision to moderate, as well as his pledge to endorse whichever candidate he liked best at the debate.

Sentiment: Negative (score ≈ -0.38)

Figure 1: An example from the BASIL dataset illustrating intra-article sentiment contrast. While surrounding sentences primarily convey neutral factual information, the biased sentence introduces a negative evaluative framing (e.g., “public ridicule”) that stands out relative to its local context.

than neutral reporting. Such methods have been applied to news articles and political text, demonstrating that sentiment-related cues are informative for identifying biased or opinionated language.

3 Data

3.1 Dataset

Sentence-level media bias detection remains relatively under-explored, and only a few publicly available datasets provide sentence-level bias annotations within news articles. In this work, we conduct experiments on two benchmark datasets: BASIL (Fan et al., 2019) and BiasedSents (Lim et al., 2020). Table 1 summarizes their key statistics.

BASIL (Fan et al., 2019) provides expert-annotated sentence-level media bias labels for 300 news articles organized into event triples reported by media outlets with different political orientations. Its multi-article, event-centered design supports analysis of bias both within and across articles, and has therefore been widely used as a benchmark for sentence-level media bias detection.

BiasedSents (Lim et al., 2020) is constructed via crowd-sourcing and contains 46 news articles covering four major events. Each sentence is annotated on a four-point bias scale, commonly converted into a binary setting. The dataset additionally provides a reference article for each event, encourag-

| Sentiment Contrast within a BiasedSents Article (CNN / Johnson) | |
|---|--|
| Sentence 1 (Context, Unbiased) | |
| <i>Republican state Rep. Dan Johnson was found dead of a single gunshot wound near Mount Washington, Bullitt County Coroner Dave Billings said.</i> | |
| Sentiment: Neutral (score ≈ 0.01) | |
| Sentence 2 (Context, Unbiased) | |
| <i>Billings said he was called to the scene around 7:30 p.m.</i> | |
| Sentiment: Neutral (score ≈ 0.00) | |
| Sentence 3 (Target, Biased) | |
| <i>She said she was staying in a living area of Louisville's Heart of Fire Church, where Johnson was pastor, when he drunkenly kissed and fondled her underneath her clothes.</i> | |
| Sentiment: Negative (score ≈ -0.52) | |

Figure 2: An example from the BiasedSents dataset illustrating intra-article sentiment contrast. While surrounding sentences primarily convey neutral factual information, the biased sentence exhibits a substantially stronger negative evaluative sentiment.

ing annotators to assess bias in a relative manner.

Since our study focuses on modeling relative sentiment contrast for bias detection, the original four-point annotation scheme in BiasedSents is not well suited for our analysis. We therefore follow prior work and adopt a binary bias formulation. Specifically, we use the preprocessed binary versions of the BASIL and BiasedSents datasets released by prior work,¹ following the binary judgment setup in (Lim et al., 2020).

Across both datasets, we observe that biased sentences often become salient through relative shifts in evaluative sentiment rather than strong sentiment polarity in isolation. As illustrated in Figure 1 and Figure 2, sentences labeled as biased introduce sentiment contrasts that stand out against their surrounding or reference context, which suggests that sentiment-based contextual contrast may provide a useful signal for sentence-level media bias detection, motivating our investigation in this work.

3.2 Dataset Analysis

Based on these sentiment observation above, we further examine article-level and local contextual sentiment deviation, and compare biased and non-biased sentences quantitatively in Table 1. Sentence-level sentiment scores are obtained using the VADER SentimentIntensityAnalyzer, which outputs polarity scores for each sentence.

¹https://github.com/yuanyuanlei-nlp/sentence_level_media_bias_naacl_2024

Figure 3 and Figure 4 illustrate the distributions of sentiment polarity and contextual deviation on the BASIL and BiasedSents datasets, respectively. For local contextual deviation, we use a fixed window of 3 surrounding sentences on each side of the target sentence.

As shown in Figure 3(a) and Figure 4(a), non-biased sentences exhibit sentiment scores concentrated around zero, reflecting a largely neutral tone, whereas biased sentences tend to skew toward more negative values. This trend is further supported by Table 1, indicating that the target sentence itself already provides informative cues for bias detection.

Additionally, clearer differences emerge when examining contextual sentiment deviation. Figure 3(b) and 4(b) show that biased sentences exhibit higher deviation from the article-level mean sentiment, suggesting that they tend to stand out emotionally within the broader narrative context of an article, which is consistent with the mean statistics reported in Table 1, where biased sentences demonstrate higher article-level deviation than their non-biased counterparts in both datasets.

The distinction becomes more pronounced at the local discourse level. As shown in Figure 3(c), biased sentences exhibit higher sentiment deviation from their surrounding context, indicating abrupt emotional shifts relative to neighboring sentences, which is consistent in the BiasedSents dataset (Figure 4), where biased sentences continue to show elevated local sentiment deviation.

Overall, these findings suggest that biased sentences are better characterized by their relative deviation from surrounding context than by sentiment polarity alone. Biased sentences tend to introduce stronger emotional contrast at both article and local levels, supporting the view that media bias emerges through contrastive evaluative framing.

4 Methods

We propose a contrastive sentiment context strategy to incorporate minimal but informative contextual signals for sentence-level media bias detection. Instead of aggregating surrounding sentences indiscriminately, our approach selects contextual sentences that exhibit strong evaluative contrast with the target sentence, based on sentiment deviation within the same article.

Let an article consist of a sequence of sentences $\{s_1, s_2, \dots, s_n\}$. For each target sentence s_i , we associate a scalar sentiment score $\text{sent}(s_i)$ obtained

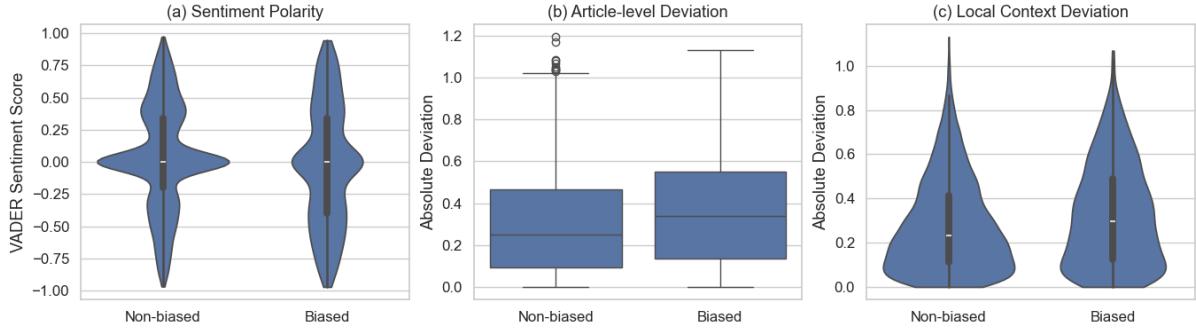


Figure 3: Distribution of sentence-level sentiment and contextual sentiment deviation on the BASIL dataset. (a) Sentiment polarity scores for both biased and non-biased sentences are largely concentrated around zero, with biased sentences exhibiting a slightly more dispersed distribution. (b) Biased sentences show higher deviation from the article-level mean sentiment. (c) Biased sentences exhibit even larger deviation from the sentiment of their local surrounding context.

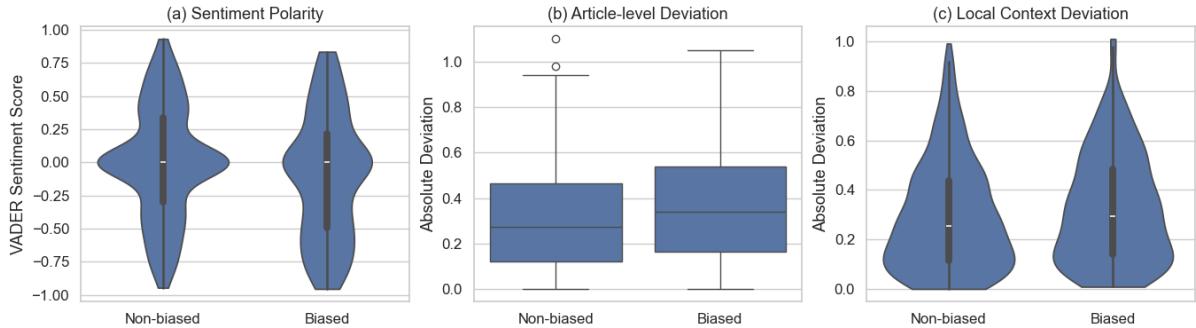


Figure 4: Distribution of sentence-level sentiment and contextual sentiment deviation on the BiasedSents dataset. (a) Sentiment polarity remains centered around zero for both classes, while biased sentences display greater dispersion. (b) Higher deviation from article-level sentiment is observed for biased sentences. (c) The deviation from local contextual sentiment is further amplified for biased sentences.

from a prior sentiment analysis step. These sentiment scores are used solely for context selection and are not treated as supervision signals during training.

To quantify evaluative contrast, we compute the absolute sentiment deviation between the target sentence s_i and every other sentence s_j in the same article:

$$\Delta_{ij} = |\text{sent}(s_j) - \text{sent}(s_i)|.$$

We then select the top- k sentences with the largest Δ_{ij} values as contrastive context:

$$\mathcal{C}_i = \text{TopK}_{j \neq i}(\Delta_{ij}).$$

To explicitly distinguish the target sentence from its contextual sentences, we introduce two special markers, [TARGET] and [CONTEXT], as learnable tokens in the model vocabulary. The token embedding matrix of the pre-trained language model is resized accordingly, allowing the model to learn

role-specific representations for target and context content.

The input for a target sentence s_i is constructed as:

$$x_i = [\text{TARGET}] s_i [\text{SEP}] [\text{CONTEXT}] \{s_j\}_{j=1}^k,$$

where $s_{j_1}, \dots, s_{j_k} \in \mathcal{C}_i$ denote the selected contextual sentences.

The sequence x_i is encoded by a pre-trained language model $f(\cdot)$, generating a contextualized representation $\mathbf{h}_i = f(x_i)$. We use the [CLS] representation and apply a linear classifier to estimate the probability of the biased class:

$$p(y_i = 1 | x_i) = \text{softmax}(\mathbf{W}\mathbf{h}_i + \mathbf{b}),$$

and train the model by minimizing the cross-entropy loss.

5 Experimental Setup

Within each experimental comparison, we keep the model architecture fixed and vary only the input

| Dataset | Bias | #Sent. | Sentiment Mean | Article Dev. | Window Dev. |
|----------------|-------------|---------------|-----------------------|---------------------|--------------------|
| BASIL | Non-biased | 6352 | 0.041 | 0.301 | 0.280 |
| BASIL | Biased | 1625 | -0.030 | 0.361 | 0.326 |
| BiasedSents | Non-biased | 552 | 0.001 | 0.311 | 0.294 |
| BiasedSents | Biased | 290 | -0.107 | 0.367 | 0.325 |

Table 1: Summary statistics of sentence-level sentiment and contextual sentiment deviation for biased and non-biased sentences across datasets. **#Sent.** denotes the number of sentences. **Sentiment Mean** denotes average sentiment polarity. **Article Dev.** and **Window Dev.** measure absolute deviation from article-level and local-context sentiment means, respectively, with higher values indicating stronger evaluative contrast.

construction strategy. We additionally experiment with different backbone encoders, but for any given set of results, the same model configuration is used across all input settings.

We compare the following five input settings:

- **Sentence-only**: a baseline setting where only the target sentence is used as input without additional context.
- **Naive window**: a fixed window of neighboring sentences surrounding the target sentence within the same article.
- **Random context**: a fixed number of sentences randomly sampled from the same article, excluding the target sentence.
- **Contrastive sentiment context**: contextual sentences are selected according to the contrastive sentiment context strategy described in Section 4, choosing sentences whose sentiment scores exhibit the largest deviation from the target sentence.

Ablation Study We conduct an ablation study to examine the effect of context size in contrastive sentiment context selection by varying the number of contextual sentences $k \in \{1, 2, 3, 4\}$. All other experimental settings are kept fixed to evaluate how the amount of sentiment-based contextual information influences bias detection performance.

Model Comparison In addition to input construction strategies, we compare two backbone encoders, BERT-BASE and RoBERTa-BASE, under the same experimental settings.

All performances are evaluated using precision, recall, and F1-score for the biased class, and results are reported as averages across validation folds.

Implementation Details All models are implemented in PYTORCH using the HUGGINGFACE TRANSFORMERS library. All models are initialized

from corresponding pre-trained language model checkpoints. Optimization is performed using AdamW with a learning rate of 2×10^{-5} and a batch size of 16, the maximum input sequence length is set to 192 tokens to keep GPU memory usage manageable. Each model is trained for 5 epochs. The naive window uses a window size of $w = 1$ (one preceding and one following sentence), the random context samples $k = 2$ sentences from the same article, and the contrastive sentiment context selects the top- $k = 2$ sentences with the largest sentiment deviation from the target sentence.

We adopt k -fold cross-validation with $k = 5$ using article-level splits via GROUPKFOLD, ensuring that sentences from the same article do not appear in both training and validation sets. We fix random seeds for Python, NumPy, and PyTorch, and enable deterministic CUDA behavior to ensure reproducibility across runs. All other hyperparameters are kept fixed to ensure a fair comparison across input configurations.

6 Results

Table 2 summarizes the performance of different input settings on BASIL and BiasedSents. Across both BERT-base and RoBERTa-base, we observe consistent trends in how different input settings affect sentence-level media bias detection.

The sentence-only setting provides a strong baseline across datasets. The sentence-only setting achieves the highest precision on BASIL with the BERT-base model, and attains the highest precision on both datasets when using RoBERTa-base, indicating that individual sentences can convey partial bias cues. However, relying on sentence-level information alone does not fully capture the broader contextual dynamics involved in biased framing.

The naive window setting generally achieves higher recall and slightly better F1 scores than random context, particularly on BiasedSents, suggest-

| Model | Input Setting | BASIL | | | BiasedSents | | |
|--------------|---------------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | | P | R | F1 | P | R | F1 |
| BERT-base | Sentence-only | 0.518 | 0.412 | 0.454 | 0.406 | 0.276 | 0.305 |
| | Naive Window | 0.504 | 0.414 | 0.451 | 0.350 | 0.282 | 0.305 |
| | Random Context | 0.515 | 0.406 | 0.451 | 0.361 | 0.279 | 0.301 |
| | Contrastive Context | 0.518 | 0.424 | 0.463 | 0.472 | 0.326 | 0.363 |
| RoBERTa-base | Sentence-only | 0.521 | 0.436 | 0.457 | 0.406 | 0.276 | 0.305 |
| | Naive Window | 0.504 | 0.426 | 0.459 | 0.350 | 0.282 | 0.306 |
| | Random Context | 0.515 | 0.406 | 0.451 | 0.362 | 0.279 | 0.301 |
| | Contrastive Context | 0.516 | 0.485 | 0.495 | 0.396 | 0.421 | 0.394 |

Table 2: Performance comparison of different input settings under two backbone encoders, BERT-base and RoBERTa-base, on BASIL and BiasedSents. All results are averaged over 5-fold cross-validation.

| Model | k | Precision | Recall | F1 |
|--------------|---|--------------|--------------|--------------|
| BERT-base | 1 | 0.561 | 0.351 | 0.428 |
| | 2 | 0.518 | 0.424 | 0.463 |
| | 3 | 0.510 | 0.433 | 0.459 |
| | 4 | 0.507 | 0.413 | 0.441 |
| RoBERTa-base | 1 | 0.498 | 0.467 | 0.475 |
| | 2 | 0.516 | 0.485 | 0.495 |
| | 3 | 0.519 | 0.464 | 0.484 |
| | 4 | 0.377 | 0.391 | 0.380 |

Table 3: Ablation study on the number of contrastive contextual sentences k for contrastive sentiment context selection on the BASIL dataset.

ing that locally adjacent sentences are more likely to share topical or discourse relevance with the target sentence, helping recover biased instances. In contrast, the random context setting tends to yield slightly higher precision but lower recall, indicating that while randomly sampled sentences are less noisy on average, they often fail to provide context that is directly informative for identifying biased framing.

In contrast, the contrastive sentiment context setting consistently outperforms other context-based approaches across both models. Although it does not always achieve the highest precision individually, particularly on BASIL with RoBERTa-base, it achieves the highest F1 scores overall. Therefore, we think that sentiment-guided contrastive selection provides a better balance between precision and recall by focusing on contextual sentences that are more directly informative for evaluative framing, rather than relying solely on proximity or random inclusion.

Overall, the results suggest that while sentence-level information remains highly informative, selec-

| Model | k | Precision | Recall | F1 |
|--------------|---|--------------|--------------|--------------|
| BERT-base | 1 | 0.351 | 0.248 | 0.285 |
| | 2 | 0.472 | 0.326 | 0.363 |
| | 3 | 0.297 | 0.194 | 0.221 |
| | 4 | 0.349 | 0.184 | 0.236 |
| RoBERTa-base | 1 | 0.390 | 0.525 | 0.440 |
| | 2 | 0.396 | 0.421 | 0.394 |
| | 3 | 0.386 | 0.424 | 0.395 |
| | 4 | 0.312 | 0.393 | 0.335 |

Table 4: Ablation study on the number of contrastive contextual sentences k for contrastive sentiment context selection on the BiasedSents dataset under different backbone encoders.

tively incorporating contextual information based on evaluative contrast offers a more effective and controlled way to complement sentence-only representations than naive or random context aggregation.

Ablation Study Tables 3 and 4 further reveal consistent yet model-specific patterns across backbone encoders. For BERT-base, $k = 2$ yields the highest F1 score on both datasets, indicating that a small amount of contrastive sentiment context offers the best balance between precision and recall. On BiasedSents, this setting outperforms other values of k across all metrics, while on BASIL, $k = 1$ achieves higher precision and $k = 3$ achieves higher recall, highlighting a clear trade-off between selectivity and coverage.

RoBERTa-base exhibits a similar but slightly more robust behavior. On BASIL, $k = 2$ again achieves the highest F1 score, driven by strong recall while maintaining competitive precision, suggesting that RoBERTa benefits more consistently from a limited amount of contrastive context.

On BiasedSents, RoBERTa shows higher recall at $k = 1$ but more balanced performance at $k = 2$ and $k = 3$, with $k = 1$ favoring coverage and larger k values gradually introducing noise.

Taken together, both models indicate that contrastive sentiment context is most effective when kept small, and that excessive contextual sentences tend to dilute the evaluative signal rather than strengthen it.

Model Comparison Comparing BERT-base and RoBERTa-base in Table 2, we observe broadly consistent trends across input settings, suggesting that the choice of backbone encoder plays a secondary role relative to how contextual information is constructed. For both models, the largest differences arise between the sentence-only and contrastive sentiment context settings, indicating that contrastive context selection interacts meaningfully with the encoder’s representation capacity.

In contrast, the naive window and random context settings yield very similar performance under both encoders, which is likely influenced by fixed random seeds and the relatively small scale of the datasets, and suggests that without principled context selection, encoder choice alone does not substantially alter model behavior.

The ablation results further reveal subtle model-specific preferences. On BiasedSents, RoBERTa-base tends to favor shorter contrastive context (e.g., $k = 1$), whereas on BASIL its performance varies more across k . Despite these differences, both encoders consistently achieve their highest F1 scores around $k = 2$, reinforcing the conclusion that a small amount of contrastive context provides the most effective trade-off between signal strength and noise across models.

Dataset Comparison Across both datasets, we observe consistent trends across input settings, suggesting that the effects of contextual strategies are largely dataset-agnostic. Overall performance is higher on BASIL, indicating that its sentence-level bias is easier to capture, likely due to expert annotations and sparser bias, which make sentence-internal cues more salient. In contrast, BiasedSents contains denser bias and relies on reference-based crowd annotations, making bias more relative in nature. This is reflected in the ablation results, where BiasedSents consistently favors shorter, strongly contrastive context, while BASIL relies more on sentence-level information with modest contextual support.

7 Discussion

We emphasize that our results are not intended to be directly compared with prior state-of-the-art systems that rely on richer supervision or complex discourse- and event-level modeling. Instead, our goal is to study media bias detection from a deliberately simplified, sentence-centric perspective, examining whether relative sentiment contrast alone provides useful signals. Unlike approaches such as (Lei et al., 2022; Lei and Huang, 2024), which employ strongly constrained event-level cross-validation and structured context modeling, our experiments aim to isolate the contribution of lightweight contextual cues. Our findings should therefore be interpreted as an exploratory analysis of simplification trade-offs rather than a competitive benchmark against complex methods.

The results suggest that sentiment-guided contrastive context is useful but limited on the two datasets studied. It is also worth noting that our experiments use two fixed model architectures and limited hyperparameter tuning, and we do not compare against more complex or task-specific bias detection models. As such, our results should be viewed as exploratory rather than conclusive, illustrating both the promise and the limitations of sentiment-based contrast as a lightweight contextual signal for sentence-level media bias detection.

8 Conclusion

In this work, we investigated sentence-level media bias detection from a minimalist perspective, treating bias as a relative phenomenon that emerges through evaluative contrast rather than explicit structural cues. We examined whether sentiment-based contrast within an article can serve as a useful contextual signal.

Regarding **RQ1**, our results show that sentence-centric models already provide good performance, suggesting that many bias cues are encoded directly in individual sentences, through lexical and stylistic choices and, in some cases, through deviations in evaluative sentiment that do not require broader contextual reasoning.

Addressing **RQ2**, the sentiment-guided contrastive context modeling is an effective strategy for sentence-level media bias detection. Although it does not always achieve the highest precision, it consistently delivers the strongest overall performance across datasets and models, demonstrating that relative sentiment contrast provides a reliable

cue for identifying biased framing. By explicitly selecting sentiment-contrastive sentences, the model is better able to focus on salient evaluative shifts that characterize media bias, rather than relying solely on local proximity or randomly sampled context.

For **RQ3**, we observe that the optimal number of contrastive contextual sentences varies across datasets, but a consistent trend emerges in favor of shorter contrastive context. In both BASIL and BiasedSents, smaller values of k reach stronger performance, while increasing k generally leads to degraded results. So sentiment-based contrast is most effective when it highlights a small number of highly informative evaluative differences, whereas incorporating additional contrastive sentences tends to dilute the signal and introduce noise.

Overall, our findings suggest that sentence-level media bias can be effectively modeled using simple, sentence-centric approaches, and that sentiment-based contrast provides a lightweight, interpretable, and complementary contextual signal. Future work may investigate combining sentiment contrast with other sentence-level cues, exploring alternative contrast selection strategies, and evaluating robustness across additional media bias datasets.

9 Reflection

Work Division The two team members contributed equally to this project. One member primarily focused on dataset analysis, experimental design, and result interpretation, while the other concentrated on model implementation, training scripts, and experimental execution. Both members jointly discussed the research questions, refined the methodology, analyzed the results, and collaborated on writing and revising the final report.

Use of AI Assistants We used an AI assistant during the research and writing process. Specifically, we used CHATGPT (GPT-5.2) to support early-stage brainstorming, drafting code prototypes and the initial experimental pipeline, debugging implementation issues, and improving the clarity and academic tone of the report. All experimental design decisions, analyses, and final writing were carried out by ourselves instead of AI assistants.

Reflecting on this process, the AI assistant was helpful for accelerating ideation and addressing low-level implementation tasks, such as boilerplate code generation and debugging. However,

its outputs occasionally contained inaccuracies or assumptions, requiring careful human verification and revision.

References

- Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. [We can detect your bias: Predicting the political ideology of news articles](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.
- Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. [Predicting factuality of reporting and bias of news media sources](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.
- Lisa Fan, Marshall White, Eva Sharma, Ruiqi Su, Pravilla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. [In plain sight: Media bias through the lens of factual reporting](#). *Preprint*, arXiv:1909.02670.
- Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. [SemEval-2019 task 4: Hyperpartisan news detection](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Yuanyuan Lei and Ruihong Huang. 2024. [Sentence-level media bias analysis with event relation graph](#). *Preprint*, arXiv:2404.01722.
- Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. 2022. [Sentence-level media bias analysis informed by discourse structures](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10040–10050, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2020. [Annotating and analyzing biased sentences in news articles using crowdsourcing](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1478–1484, Marseille, France. European Language Resources Association.
- Iffat Maab, Edison Marrese-Taylor, and Yutaka Matsuo. 2023. [Target-aware contextual political bias detection in news](#). *Preprint*, arXiv:2310.01138.
- Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. [Linguistic models for analyzing and detecting biased language](#). In *Annual Meeting of the Association for Computational Linguistics*.