# Target-Driven Contexts in Detecting Informational Bias

Iffat Maab[1], Edison Marrese-Taylor[1, 2], Yutaka Matsuo[1]
Graduate School of Engineering, The University of Tokyo[1]
National Institute of Advanced Industrial Science and Technology[2]
{iffatmaab, emarrese, matsuo}@weblab.t.u-tokyo.ac.jp

## Abstract

Media bias detection requires comprehensive integration of information derived from multiple news sources. Sentence-level political bias detection in news is no exception, and has proven to be a challenging task that requires an understanding of bias in consideration of the context. Inspired by the fact that humans exhibit varying degrees of writing styles, resulting in a diverse range of statements with different local and global contexts, previous work in media bias detection has proposed augmentation techniques to exploit this fact. Despite their success, we observe that these techniques introduce noise by overgeneralizing bias context boundaries, which hinders performance. To alleviate this issue, we propose techniques to more carefully search for context using a bias-sensitive, target-aware approach for data augmentation. Our approach outperforms previous methods significantly when combined with pre-trained models such as BERT.

## 1 Introduction

Biased media have the potential to sway readers in potentially detrimental paths. Therefore, it is crucial to unveil the true nature of media bias. We think bias detection is important as a proxy or mechanism to assess the quality of information in news media. As stated by [1], there is no problem with the existence of narratives in substandard journalism, rather poor professionalism. A study by [2] suggests that indeed media has a sizable political impact on voting, where for example [3] found significant effect of exposure to Fox News in increased turnout to the polls.

Bias in news from different aspects has been studied in the past, where for example [4] and [5] created news quality corpus of 561 articles and study how various news constituents characterize the quality of editorial articles. While these works are highly relevant to the bias problem, they did not specifically or directly target at the issue.

Foundational work in political bias was performed by [6], who released a human-annotated dataset named Bias Annotation Spans on the Informational Level (BASIL), containing 300 fine-grained bias annotations. Concretely, political bias is identified at the sentence-level, where spans are annotated and a target (the main entity) is identified, in addition to a few other labels. Significantly, BASIL stands as the first dataset to be annotated with different types of bias. **Informational** bias, which depends broadly on the context of the sentence [7] and arises from manipulation of information or selective presentation of content in a factual way, e.g., use of quotes, to evoke specific reader's emotions towards news entities [6, 8], and **lexical** bias, which stems from the choice of specific words or linguistic phrases that influence the interpretation of a subject, and perpetuate the understanding of information [9, 10, 11] are present in BASIL. To the best of our knowledge, BASIL is the first dataset that annotates informational bias together with specific targets.

With the release of BASIL, work on political bias detection has mostly focused on informational bias, with a strong emphasis on informational context within and across news media articles, as informational bias is highly content-dependent. In the seminal work, [8] feed the whole document/article as context for sentence-level bias classification. Though this approach worked relatively well in practice, using long documents in this context brings considerable noise, redundancy and can increase vocabulary size, which can ultimately decrease the performance of the classifier as evidenced by previous work [12]. Moreover, as shown by [13], detecting bias at article level remains even more challenging and difficult task.

In light of this issue, several works have recently focused on introducing more specific contextual information to perform classification [14, 8, 12]. While the aforemen-

| Source | Target | Index | Sentence | Bias |
|---|---|---|---|---|
| FOX | Obama Campaign | 0 | President Obama health care plan treats the treasured entitlement like a piggy bank, while the Romney-Ryan plan preserves it. | Inf |
| HPO | | 4 | If any person in this entire debate has blood on their hands in regard to Medicare, it's Barack Obama. | Inf |
| NYT | | 4 | Now when you need it, Obama has cut $716 billion from Medicare. | Inf |

Table 1: Bias sentences extracted from event 0 of BASIL with three news media sources, FOX (0fox; source:fox, event:0), HPO (0hpo; source:hpo, event:0), and NYT (0nyt; source:nyt, event:0), showing a single event can exhibit similar targets and bias types to manifest event-based target aware context.

tioned approaches have resulted in improved performance, we think their applicability is limited. On one hand, articles in BASIL have no overall bias label, instead each sentence is labeled as evidence of a certain kind of bias or as a neutral statement, suggesting that these should be treated separately when detecting different kinds of bias. Previous studies [15, 16] have already shown that on document-level classification, paragraphs can belong to multiple categories, which [13], also observed on BASIL, where paragraphs belong to either informational bias, lexical bias or no bias spans. Furthermore, as highlighted by [13], by mixing contexts of informational and lexical bias, it becomes difficult for the model to distinguish and predict different type of bias, which may result in lower model performance.

In this work, we provide a framework to generate more consistent and similar bias contexts to improve performance. As shown in Table 1, each instance of annotated bias span also identifies the "target", i.e., the main entity or topic of the sentence that is also annotated in BASIL. Using this information, our key insight is to create event-level contexts that are target-aware and also sensitive to the bias label. For example, for the target "Obama Campaign", sentences from three different news sources are combined to form a single contextual example for informational bias classification, as highlighted in light gray. Inspired by ideas from modeling context in informational bias detection [8, 13, 12], our approach is able to augment examples with richer contexts and less noise. Following recent work [17, 8, 12], we tackle a bias detection task of INF/OTH using data from BASIL.

Through extensive experimentation, we demonstrate the effectiveness of our approach by obtaining state-of-the-art performance on all of our studied tasks. In addition, our holistic view on bias enables us to unveil inconsistent terminologies used for contextual information of BASIL, therefore we gather such contexts to improve clarity and uniformity, and to avoid previous work problems as indicated in our comparison with the state-of-the-art.

## 2 Related Work

Media bias has been scrutinized often with nuanced variations and under different contexts through diverse terminologies. [18] proposed attention-based model to capture high level contexts of news articles including title, link structure, and news information using both textual content and network structure to leverage cues from multiple views. Contextualized representations of sentences for better understanding of documents are studied using numerous pre-trained language models [14, 10]. Inspired from [14], [8] work on BASIL to propose several context inclusive models on article and event context, and use three BiLSTMs for encoding FOX, HPO, NYT news documents as triplets. Building upon existing study of [8], [12] use multi-level graph attention networks for bias detection by MultiCTX model that use contrastive learning from sentence embeddings to discriminate target sentences. Another recent study on BASIL [19] built distillation models on top of RoBERTa for informational bias classification and explore different types of local and global discourse structures. Similarly, article-level bias classifiers [13] use second order bias features of BASIL to manipulate context information using uncased BERT. Using BASIL, BERT by [20] remain as a major baseline model in majority of previous studies [8, 12]. [13] find that fine-tuned BERT has a strong efficacy and use it to reimplement [6] results. In light of the findings, our proposed approach also utilize BERT [20].

## 3 Proposed Approach

### 3.1 Bias-Aware Neighborhood Context

Previous work has shown that phrases surrounding a sentence annotated with bias can be used as local context to perform bias classification, and that this local context can contribute to the ability of models to identify and label types of bias. However, by ignoring the nature of these

| | Target | Sentences | | | Target-aware examples | | | | Total |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Article-level | | | Event-level | |
| | | FOX | HPO | NYT | FOX | HPO | NYT | | |
| 18 | Benjamin Netanyahu | 1 | - | - | 1 | - | - | - | 1 |
| | Barack Obama | 5 | 1 | - | 10 | 1 | - | 5 (fox × hpo) | 16 |
| | Secure America Now | - | 2 | 2 | - | 1 | 1 | 4 (hpo × nyt) | 6 |
| **Total** | | | | | within Art. = 14 | | | 9 | **23** |
| 22 | Hillary Clinton | 5 | - | 3 | 10 | - | 3 | 15 (fox × nyt) | 28 |
| | Barack Obama | 2 | 2 | - | 1 | 1 | - | 4 (fox × hpo) | 6 |
| | Nancy Pelosi | 1 | - | - | 1 | - | - | - | 1 |
| **Total** | | | | | within Art. = 16 | | | 19 | **35** |

Table 2: Detail of the number of contextualized instances obtained by applying our proposed ABTA and EBTA to a set of the original examples from BASIL, in this case taken from events (E) 18 and 22, for the case of informational bias.

sentences, existing approaches that utilize neighborhood context [8, 12] can run into problems by introducing ambiguous content, for example when adding sentences that are annotated with the opposite bias. As shown by [8], this can also lead to massive data leakage problems across train and test sets.

To account for the disparity in how different bias contexts are overlooked in previous work, in this paper, we propose to care for the bias label of neighboring sentences, advancing to generate Bias-Aware Neighborhood Contexts (BANC), and adding neighboring sentences to the model input as long as they have a related bias label. See Appendix B for the detailed description.

## 3.2 Target-Aware Context

While our neighboring approach helps identify local context relevant for bias classification, we believe that global context, either at the article or event levels, can also be exploited. To that end, we note that BASIL contains annotations that also identify the "target" of a given sentence where informational bias is present. This "target" label refers to the main entity or topic of the sentence that is annotated, with some of the most prominent targets in BASIL being entities or people that lie at the core of news reports, such as Donald Trump, Romney Campaign, Secure America Now, among others.

We further note that although the frequency of appearance of a given "target" varies substantially, as long as we keep the annotated label constant (e.g., informational), the context remains the same. This motivated us to gather all surrounding linguistic cues pertaining to a specific target at both the article-level and event-level. Concretely, we create target-aware contextual information by making use of all possible combination spans having the same bias and target, and propose article-based target-aware (ABTA) and event-based target-aware (EBTA) contexts, which we explain below.

As show in Table 2, using ABTA context, for instance, the target "Barack Obama" which has 5 sentences annotated with informational bias in the FOX article and 1 in HPO, generates all possible combinations of two sentences within FOX giving us 10 contextualized examples, and 1 same example in HPO because this article has only one sentence, respectively. Note that possible combinations of sentences within articles are combined in groups of two only, which we do to emulate the natural distribution of occurrence of sentences with the same bias and same "target".

EBTA contexts shown in the "Event-level" column in Table 2, are computed for common targets across articles, for instance, the same target "Barack Obama" with informational bias appear across FOX and HPO with 5 and 1 sentences, therefore all unique possible combinations in groups of two generates 5 new contextualized examples across the two aforementioned articles. Finally, following the example in the table for "Barack Obama", the combined contexts of ABTA and EBTA give us a total of $10+1+5=16$ contextualized informational bias examples for a single target.

Because of the way in which we combine sentences, it is evident that our approach is significant in providing contextualized examples for infrequent targets as well, therefore also contributing towards mitigating imbalanced bias distribution and skewed nature of "targets" as observed in BASIL articles [13].

Using our target-aware techniques, we observe a four-fold increase of examples for informational bias detection (1,221 original BASIL sentences v/s 4,987 contextualized examples). Please see Appendix C for the most frequent "targets" in BASIL. Finally, based on successful results reported by previous work [21, 17], we additionally use a backtranslation approach to generate more data, which we apply to our contextualized samples using Spanish as a pivot language.

By incorporating multiple viewpoints in our neighborhood and target-aware contexts, we facilitate our model in providing a broad and inherent semantics of biased targets

| Model | INF / OTH | | | |
|---|---|---|---|---|
| | Acc. | P | R | INF F1 |
| **Neighborhood Context** | | | | |
| SSC-5 [8] | - | 41.90 | 36.16 | 38.19 |
| SSC-10 [8] | - | 43.84 | 34.88 | 38.22 |
| WinSSC-5 [8] | - | 42.28 | 36.94 | 38.67 |
| WinSSC-10 [8] | - | 43.20 | 35.12 | 37.44 |
| RoBERTa [8] | - | 43.12 | 41.29 | 42.16 |
| MultiCTX [12] | - | 47.18 | 44.01 | 45.53 |
| BERT + BT [17] | 83.86 | **51.22** | 46.32 | 50.70 |
| BERT + BANC (ours) | 83.72 | 49.07 | 45.32 | 48.90 |
| BERT + BANC + BT (ours) | **85.31** | 50.08 | **48.12** | **52.07** |
| **Article Context** | | | | |
| WinSSC [12] | - | 41.47 | 34.37 | 37.58 |
| ArtCIM [8] | - | 38.81 | 47.78 | 42.80 |
| **Event Context** | | | | |
| EvCIM [8] | - | 39.72 | 49.60 | 44.10 |
| EvCIM [12] | - | 47.07 | 44.64 | 45.81 |
| BERT [13] | - | 58.62 | 32.08 | 41.46 |
| RoBERTa [19] | - | 43.53 | 49.84 | 46.47 |
| MultiCTX [12] | - | 47.78 | 44.50 | 46.08 |
| BERT + ABTA + EBTA (ours) | 84.36 | 52.78 | 47.74 | 52.91 |
| BERT + ABTA + EBTA BT (ours) | **86.05** | **54.10** | **49.82** | **54.46** |
| BERT + BANC + ABTA + EBTA (ours) | 84.90 | 55.60 | **53.93** | 56.88 |
| BERT + BANC + ABTA + EBTA + BT (ours) | **86.40** | **59.22** | 53.12 | **58.15** |

Table 3: Comparison of our approach with previous work, separated by usage of context. We report average results of three runs with different random seeds. In the Table, Acc, P, and R stand for Accuracy, Precision and Recall respectively. BT denotes the augmentation approach from [17], who are also the only authors to report accuracy.

to manifest variations in bias representations. Our experiments will further demonstrate the impact of proposed context in different training settings.

## 4 Results and Experiments

See Appendix A for experimental setup and implementation details.

**Baselines** We consider multiple contextual models that address the detection of informational bias, for example, SSC (Sequential Sentence Classification) [22] and its variant WinSCC (windowed Sequential Sentence Classification) [8], RoBERTa, ArtCIM for target sentences within an article, and EvCIM for triplets of articles covering the same event [8, 12]. [12] further proposed MultiCTX model and reproduce the results using WinSCC and EvCIM for informational bias detection. We also compare against the fine-tuned RoBERTa model [19], as well as BERT [17, 13, 20, 6].

We compare our model with various baselines against most studied INF/OTH bias task of BASIL using contextual information as indicated by prior work [17]. Based on our comprehensive analysis on how prior studies use different contexts on BASIL, we align similar contexts of our proposed method to allow meaningful comparisons as shown in the Table 3, using three corresponding sections. To compare with previous work where only within article context is used, we concretely utilize our top performing models for comparison, i.e., BERT combined with 100% BANC (BERT + BANC), and with backtranslation (BERT + BANC + BT). Similarly, prior work using event contexts are compared with our BERT model trained on 100% target-aware (BERT + ABTA + EBTA), and with backtranslation (BERT + ABTA + EBTA + BT), respectively. Since MultiCTX by [12] uses multi-contrast learning of both article and event contexts, we compare and use our best BERT model with fusion of both proposed context techniques (BERT + BANC + ABTA + EBTA), and with backtranslation (BERT + BANC + ABTA + EBTA + BT), which in essence is our final model. Based on our results, and supporting findings of our ablation study, both BANC and target-aware (ABTA & EBTA) hold significance in our approach, however target-aware contexts contributes more than BANC parallel to previous findings [12]. Our approach outperforms previous work significantly, obtaining an F1-score of 58.15 in INF label.

## 5 Conclusion

We study a challenging and significant task of detecting misinformation and shed light on bias prevalence in news media. Our work focus on incorporating bias sensitive (BANC) and target-aware contexts (ABTA & EBTA) for sentence-level bias detection tasks. Our model encompass the process by which individuals acquire new knowledge in real-world settings, i.e., gathering the associated type of bias from common news media targets covering the same event coupled with experiences, and subsequently utilizing such contexts to make predictions about unfamiliar aspects. Our model concretely outperforms classification performance of strong baselines, and we find that the best performance is achieved when target-aware contexts are combined with BANC, and our methodological standpoint in using small-augmented data of frequent targets suggests that our model is better at recognising bias in media.

## Acknowledgements

The authors wish to express gratitude to the funding organisation as this work has been supported by the Mohammed bin Salman Center for Future Science and Technology for Saudi-Japan Vision 2030 at The University of Tokyo (MbSC2030).

## References

[1] Unesco. Journalism, 'fake news' & disinformation: handbook for journalism education and training. **UNESCO**, 2018.

[2] Tim Groseclose and Jeffrey Milyo. A measure of media bias. **The quarterly journal of economics**, Vol. 120, No. 4, pp. 1191–1237, 2005.

[3] Stefano DellaVigna and Ethan Kaplan. The fox news effect: Media bias and voting. **The Quarterly Journal of Economics**, Vol. 122, No. 3, pp. 1187–1234, 2007.

[4] Wei-Fan Chen, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. Learning to flip the bias of news headlines. In **Proceedings of the 11th International conference on natural language generation**, pp. 79–88, 2018.

[5] Ioannis Arapakis, Filipa Peleja, Barla Berkant, and Joao Magalhaes. Linguistic benchmarks of online news article quality. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1893–1902, 2016.

[6] Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 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)**, pp. 6343–6349, Hong Kong, China, November 2019. Association for Computational Linguistics.

[7] Shijia Guo and Kenny Q. Zhu. Modeling multi-level context for informational bias detection by contrastive learning and sentential graph network, 2022.

[8] Esther van den Berg and Katja Markert. Context in informational bias detection. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 6315–6326, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[9] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1650–1659, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[10] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political ideology detection using recursive neural networks. In **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1113–1122, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[11] Christoph Hube and Besnik Fetahu. Neural based statement classification for biased language. In **Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining**. ACM, jan 2019.

[12] Shijia Guo and Kenny Q Zhu. Modeling multi-level context for informational bias detection by contrastive learning and sentential graph network. **arXiv preprint arXiv:2201.10376**, 2022.

[13] Wei-Fan Chen, Khalid Al-Khatib, Benno Stein, and Henning Wachsmuth. Detecting media bias in news articles using gaussian bias distributions. **arXiv preprint arXiv:2010.10649**, 2020.

[14] Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S Weld. Pretrained language models for sequential sentence classification. **arXiv preprint arXiv:1909.04054**, 2019.

[15] Guozheng Rao, Weihang Huang, Zhiyong Feng, and Qiong Cong. Lstm with sentence representations for document-level sentiment classification. **Neurocomputing**, Vol. 308, pp. 49–57, 2018.

[16] Abinash Tripathy, Abhishek Anand, and Santanu Kumar Rath. Document-level sentiment classification using hybrid machine learning approach. **Knowledge and Information Systems**, Vol. 53, pp. 805–831, 2017.

[17] Iffat Maab, Edison Marrese-Taylor, and Yutaka Matsuo. An effective approach for informational and lexical bias detection. In **Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)**, pp. 66–77, 2023.

[18] Vivek Kulkarni, Junting Ye, Steven Skiena, and William Yang Wang. Multi-view models for political ideology detection of news articles. **arXiv preprint arXiv:1809.03485**, 2018.

[19] Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. Sentence-level media bias analysis informed by discourse structures. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 10040–10050, 2022.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

[21] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. **arXiv preprint arXiv:1309.4168**, 2013.

[22] Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. Pretrained language models for sequential sentence classification. 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)**. Association for Computational Linguistics, 2019.

## A Experimental Setup

To streamline the comparison with prior work [8, 12], we use a 10-fold cross-validation setting where bias-aware neighborhood and event-based target aware contexts never appear at the same time in non-overlapping train-val-test split sets of 80-10-10, respectively. Average performance of our model using three seed runs is reported in all our experiments.

For the sentence-level bias detection, we perform of informational bias i.e., INF/OTH bias task. Inspired by [17], for INF/OTH bias task we combine BANC, EBTA and ABTA with backtranslation of informational bias samples.

We refer to the original set of examples in BASIL, without augmentation as "regular". We do not perform any augmentation techniques for the testing examples . Furthermore, to examine the effectiveness of our proposed components in ablation studies, regular BASIL examples [6] are augmented with BANCand target-aware contexts in fractions of 10%, 20%, 30%, 40%, 50%, 100%, and 100% with BT (additional backtranslated examples).

**Implementation Details** We use the PyTorch to implement our models, borrowing from HuggingFace, our classifiers are based on BERT-base [20], and all our models are trained with $5 \times 10^{-5}$ as learning rate, 32 as batch size, and 15 as a maximum epoch count. We utilize a server with an NVIDIA V-100 GPU for our experiments.

| Index | Position | Sentence | Bias |
|---|---|---|---|
| 2 | Neighbor | The fact is that every day that passes, Iran gets closer and closer to nuclear arms, Mr. Netanyahu is shown saying. | - |
| 3 | Target | For dramatic effect, a soundtrack fit for an episodic drama like Homeland plays as the prime minister continues. | INF |
| 4 | Neighbor | The world tells Israel, Wait. There's still time. | - |
| 5 | Neighbor | And I say wait for what? | - |
| 6 | Neighbor | Wait until when? | - |

Table 4: An article of New York Times section extracted from BASIL showing bias-aware neighborhood context of informational bias in blue.

## B Bias-Aware Neighborhood Context

Table 4 shows an example of how this procedure works. Since, our approach is bias-sensitive, to generate a BANC for informational bias classification, we combine sentences with indices 2, 3 and 4 as highlighted in blue. According to the same principle, for cases where the first sentence of an article has bias, next sentence is checked and combined, whereas in the event where it is last sentence, former sentence gets checked and successively combined.

## C Most Frequent Targets

Table 5 shows a detailed explanation on target-aware context generation for the most frequent "targets" in BASIL.

| Target | Target Aware Context | |
|---|---|---|
| | Sentences | Possible Combinations |
| Donald Trump | 340 | 2767 (Inf: 2386, Lex: 381) |
| Barack Obama | 119 | 619 (Inf: 479, Lex: 140) |
| Barack Obama* | 156 | 870 (Inf: 705, Lex: 165) |
| Hillary Clinton | 62 | 327 (Inf: 292, Lex: 35) |
| Democratic Lawmakers | 36 | 119 (Inf: 97, Lex: 22) |
| Joe Biden | 32 | 325 (Inf: 241, Lex: 84) |
| Paul Ryan | 25 | 122 (Inf: 97, Lex: 25) |

Table 5: Most frequent bias targets in BASIL across events and their possible combinations using target-aware context. Barack Obama* includes three similar targets: Barack Obama, Obama's administration, Sasha and Malia Obama with 119, 21, and 16 bias sentences.