Political Bias detection using transformer-based models

Tatiana Getling

January 7, 2025

1 Introduction

In today's digital age, the issue of political bias detection has become increasingly important. Media coverage plays a crucial role in shaping public perception of events and influencing both individual and collective opinions. While news articles are intended to provide objective and transparent information, they frequently reflect the author's biases—consciously or otherwise—which can sway readers' interpretations of events. Such biases are particularly evident in coverage of controversial topics, often characterized by inflammatory language, one-sided statements, or inaccuracies [5]. Phenomena like the Hostile Media Effect, i.e. describing the tendency to perceive media coverage of an issue as biased against one's views, also plays a role, making it hard to objectively determine whether and how an article or clip is biased [12].

Detecting and mitigating political bias is a challenging task, even for human experts. Bias can manifest subtly in the form of word choice, framing, or selective omission of facts, requiring careful contextual analysis to identify. The complexity of this task is further compounded when attempting to automate the process. Despite significant advancements in natural language processing (NLP) and machine learning, the development of reliable models for bias detection remains an open research area. One of the primary challenges is the lack of large, high-quality ground-truth datasets that adequately represent the nuanced nature of bias in political texts.

Recent advancements in transformer-based models, such as BERT and its variants, have shown promising results in numerous NLP tasks, offering state-of-the-art performance in text classification, sentiment analysis, and question answering. These models, with their ability to capture contextual and semantic relationships, hold potential for addressing the complexities of political bias detection. However, their effectiveness in this specific domain has yet to be thoroughly evaluated. This project aims to contribute to the field of political bias detection by exploring the application of transformer-based models.

2 Related Works

Existing research on bias detection can be broadly categorized based on the techniques employed in the detection process. A systematic literature review by [17] highlights three main categories of approaches: traditional natural language processing (NLP) techniques, machine learning (ML) methods, and graph-based approaches.

Traditional NLP techniques, further divided into count-based and embedding-based approaches, have been foundational in detecting bias, though their effectiveness can be limited due to the complexity of modern bias manifestations. These methods are frequently used as a baseline when introducing a new dataset due to their explainability and proven effectiveness [17]. More recently, machine learning methods have gained prominence, with further subdivisions into transformer-based models (e.g., BERT, RoBERTa) and non-transformer-based models, each offering distinct advantages in capturing the nuances of media bias. Additionally, non-neural network ML techniques, such as support vector machines and decision trees, continue to be explored for their robustness in certain bias detection tasks. Graph-based approaches, such as those focusing on discourse structures or topic modeling, have also emerged as effective tools for analyzing the relationships between entities and bias.

Machine learning approaches, particularly those leveraging transformer-based models, dominate current research. [6] employ domain-adapted versions of BERT and RoBERTa (DA-BERT, DA-RoBERTa, and DA-BART), pre-trained on media bias datasets, to enhance bias classification performance. Their models, evaluated on the BABE dataset, exhibit significant improvements over general-purpose baselines.[18] investigate the role of context in informational bias detection using the BASIL dataset. They compare various models incorporating direct textual context (e.g., WinSSC), article-level context (ArtCIM), and event-level context (EvCIM), and show that direct textual context and domain context are difficult to integrate in a way that boosts performance beyond the RoBERTa baseline. They also find that context-inclusive model outperforms RoBERTa significantly when using event context (EvCIM). They further perform error analysis that shows that both models are better at recognizing bias in quotes, while EvCIM performs better than the baseline on longer sentences and sentences from politically centrist articles. [11] propose a framework for de-biasing news articles through a four-phase pipeline: detection, recognition, masking, and de-biasing of a biased text. They use the MBIC dataset and compare the performance on several models (logistic regression + TFIDF, random forest + TFIDF, gradient boosting machine + TFIDF, logistic regression + ELMO, multi-layer perceptron + ELMO, BERT-base, RoBERTa-base, DistilBERT). The DistilBERT model is used for the additional fine-tuning as showing the best performance. This framework not only detects but also actively mitigates bias, representing a novel contribution to actionable bias reduction. [16] identify bias by word choice with the Multi-Task learning approach based on DistilBERT using in-domain and cross-domain datasets for learning and the BABE dataset for the evaluation. Although their multitask learning model surpasses baselines in some cases, the approach shows limited improvement in bias classification tasks, highlighting the inherent challenges of this problem domain.

2.1 Datasets

The field of bias detection research lacks high-quality datasets. Most available datasets are predominantly in English, which limits the generalizability of findings. Furthermore, most datasets are either small or rely on crowdsourcing for bias attention, which can be a difficult task even for domain experts. The perception of bias is affected by many individual factors, including topic knowledge, political ideology, age and education [13].

These factors contribute to low inter-rater agreement, making consistent annotation a persistent challenge.

One of the most frequently used datasets for the bias detection is BASIL, introduced by [3], which contains 100 sets of triplets of news articles discussing the same event from news outlets with different political leanings. The articles are annotated by two annotators (for each article) with 1727 bias spans, containing labels for both informational and lexical bias at the sentence and document levels. BASIL has been widely utilized in research, including works by [18], [10], [4], [2], [8], [7].

The MBIC (Media Bias Including Characteristics) dataset, created by [15], contains 1700 statements that represent various media bias instances. Each statement is annotated at the word and sentence levels by ten annotators. Unlike most crowdsourced datasets, MBIC is the first dataset to include the annotators characteristics and their individual background, offering valuable metadata for understanding annotation biases.

The BiasedSents dataset, introduced by [9], is also annotated using crowdsourcing, and consists of 966 sentences from 46 English-language news articles covering 4 different events and with labels provided on the sentence level.

Another notable resource is the BABE (Bias Annotations by Experts) dataset introduced by [14]. The dataset focuses on media bias instances annotated by domain experts. Unlike other datasets that mostly use crowdsourcing for annotation, BABE emphasizes reliability and quality of annotations. It has been applied in studies analyzing nuanced forms of media bias, particularly in cases where expert judgment is crucial.

[1] created a dataset containing 34,737 articles that were manually annotated for political ideology using labels from AllSides.com. In AllSides, these annotations are made as a result of a rigorous process that involves blind bias surveys, editorial reviews, third-party analysis, independent reviews, and community feedback [1]. This dataset is particularly useful for political ideology detection, offering a structured framework for understanding media bias through a stance-detection lens. It incorporates metadata such as article sources and alignment labels.

3 Methodology

3.1 Dataset Description

The primary dataset used in this study is the MBIC (Media Bias Including Characteristics) dataset created by Spinde et al. [15]. This study leverages MBIC due to its well-curated content and relevance to the domain of political bias detection.

In addition to MBIC, we utilized the dataset created by Baly et al. [1], which includes textual data from multiple news outlets with annotated political bias. While this dataset is designed for the political ideology prediction and is annotated accordingly, it contains a larger amount of data from the domain with respect to MBIC, which is why we used it for the n-grams extraction in order to enable a comparative analysis with the n-grams derived from MBIC. By incorporating this additional dataset, we aim to assess the generalizability and domain-specific patterns of n-gram usage across different data sources.

3.2 Data Preprocessing

The preprocessing pipeline was designed to standardize the data and prepare it for efficient modeling. The data preprocessing included lowercasing the data to ensure uniformity, removing any special character and stopwords from the NLTK stopword list, as well as tokenizing the data to facilitate further text analysis. Furthermore, the data was split into training and test sets using the 80/20 split.

3.3 Model Selection

In our project we compare two transformer-based models: a baseline BERT model and a DistilBERT model that have been proven to be effective in political bias detection in several studies.

Due to the limited computational resources available for this study, we were unable to extensively experiment with different configurations of the BERT model. While BERT's larger architecture often yields state-of-the-art results, its training and fine-tuning require significant computational power, particularly for hyperparameter optimization and exploring alternative configurations such as batch size, learning rate, and sequence length.

As a result, we opted to use default parameter settings and focused our efforts on fine-tuning DistilBERT, a lighter version of BERT, which is specifically designed for resource-constrained environments. Although this approach allowed us to complete the experiments within the available computational budget, it may have restricted the potential performance improvements achievable with the full BERT model.

Future work could address this limitation by leveraging high-performance computing resources to conduct comprehensive hyperparameter tuning and explore other variations of BERT, such as RoBERTa or BERT-large, which could potentially lead to more robust results.

To further evaluate the generalizability of the models and the potential need for fine-tuning on larger datasets, we incorporated n-gram analysis on two distinct datasets. This approach aimed to compare the performance and feature importance of the models trained on the smaller dataset against those trained or tested on the larger one. By examining the impact of n-gram features, we sought to determine whether the smaller dataset captured sufficient linguistic patterns for reliable bias detection or if fine-tuning the model on the larger dataset would yield improved results. This analysis provided insights into the transferability of features across datasets and informed decisions regarding the scope and scale of future training efforts.

4 Experimental Setup

4.1 Software Environment

The experiments were conducted in a Python environment utilizing popular libraries for natural language processing and machine learning. Key software components included libraries, such as Transformers (Hugging Face) for BERT-based model implementation, Scikit-learn for dataset splitting and evaluation metrics, Pandas and NumPy for data manipulation and NLTK for stopword removal and n-gram processing. Furthermore, the experiments were conducted on Google Colab platform, due to its ease of use and the

provision of high-performance computing resources, enabling efficient training and testing of the models.

4.2 Model Setup

The baseline BERT model was trained over three epochs with a batch size of 8 and a learning rate of 5e-5. The AdamW optimizer was employed to minimize the cross-entropy loss function. For the DistilBERT model we compared the pre-trained model trained over 3 to 5 epochs, with different learning rate and batch size parameters.

5 Results

5.1 N-gram Analysis

The performance of n-gram analysis was evaluated using frequency counts and their predictive utility as features in classification tasks.

For the initial comparison, we have compared the top 15 n-grams extracted from both datasets.

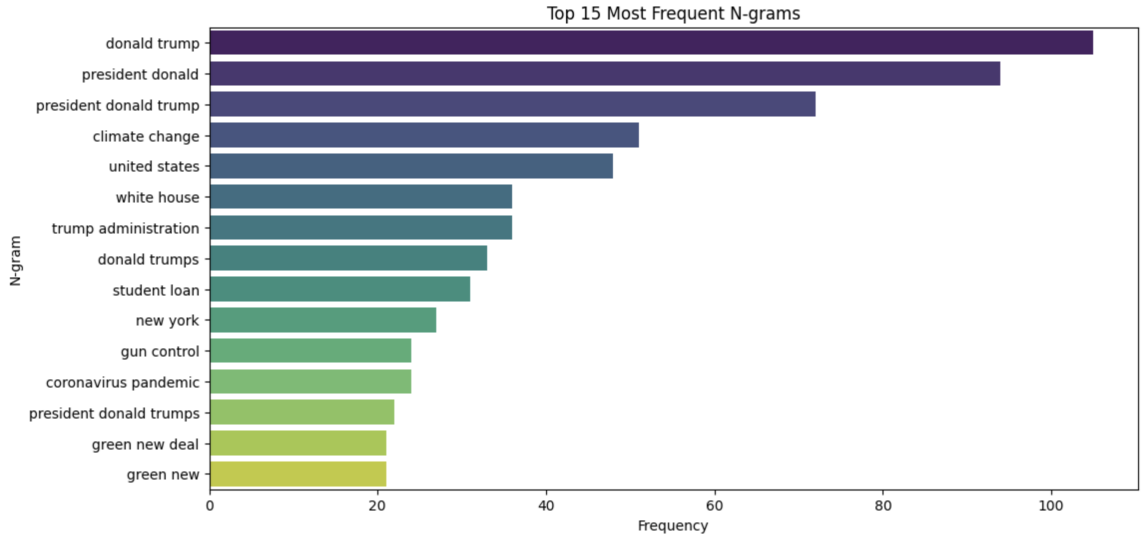

Figure 1: Top 15 N-grams from the MBIC dataset.

We can see that the n-grams differ significantly between the two datasets. In fact, out of the 15 most frequent n-grams, only one appears in both results:

These differences indicate that the extraction of the n-grams from the MBIC dataset might not be enough for a correct representation of the domain, so if we were to include n-grams as a feature for our model, we would have to use the n-grams from a larger dataset in order to get an accurate representation of the domain.

5.2 Evaluation Metrics

To assess the performance of the models, we utilized a set of standard evaluation metrics commonly used in classification tasks: accuracy, precision, recall and F1-score.

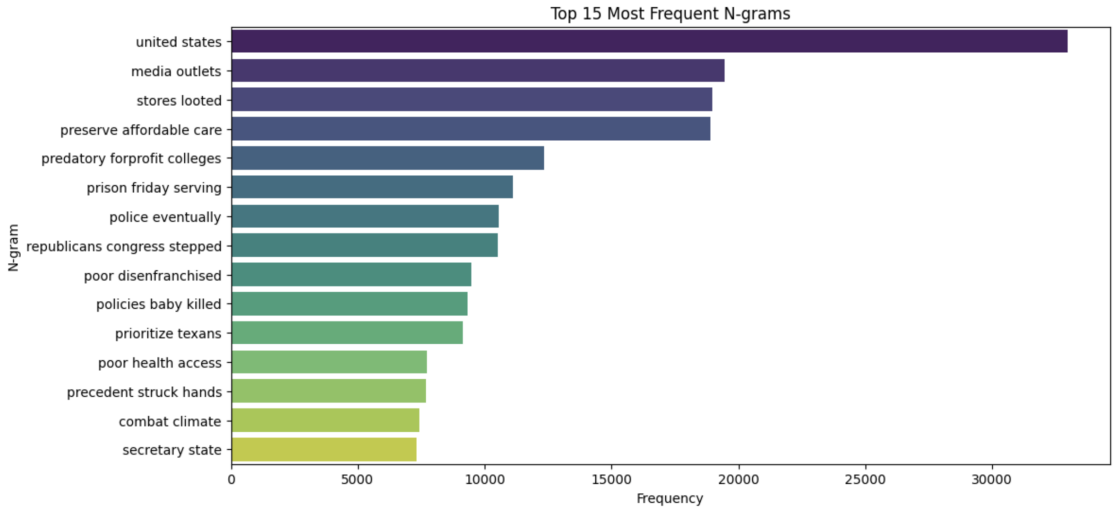

Figure 2: Top 15 N-grams from the Baly dataset.

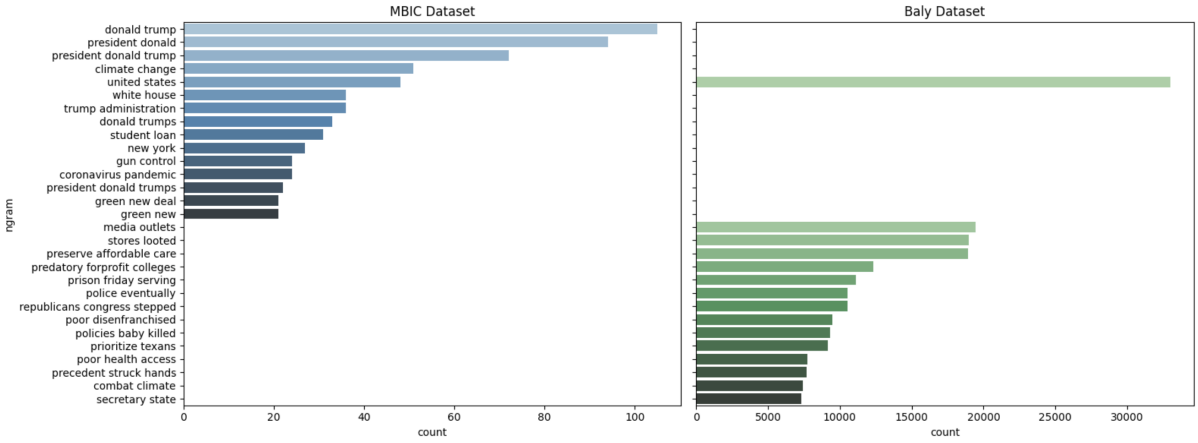

Figure 3: Top N-grams comparison.

The baseline experiment utilized the BERT model, with the 5e-5 learning rate, batch size 8 abd 3 epochs. The model achieved the following results:

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
BERT	0.7474	0.7359	0.7474	0.728

Table 1: Baseline BERT model performance

These results demonstrate a strong performance of the BERT model in capturing the nuances of political bias.

Given the computational limitations of training BERT, the lightweight DistilBERT model was explored as an alternative. The initial DistilBERT tests were performed with the same parameter as the baseline BERT model: 3 epochs, batch size of 8 and the learning rate 5e-5. The initial model achieved the following scores, showing that the default setting yield a significantly lower performance compared to the BERT model, thus indicating the need for further tuning:

In order to improve the model's performance, we peformed the hyperparameter tuning using Optuna. The optimization setup included 3- and 5-epoch training for each trial

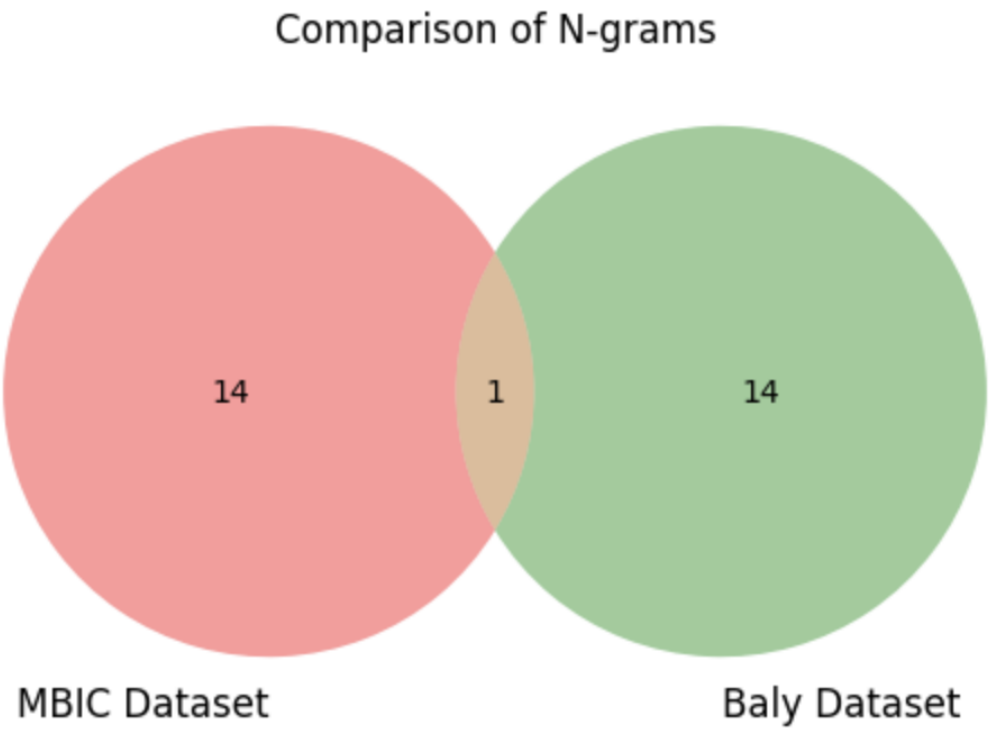

Figure 4: Vienn diagram of the N-grams between the two datasets.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
DistilBERT	0.3071	0.0943	0.30716	0.14436

Table 2: First DistilBERT model performance

(using 5 and 10 trials), batch sizes of 8, 16, and 32, learning rate from 1e-5 to 1e-3, dropout values from 0.1 to 0.5 and attention dropout from 0.1 to 0.5. The best performance was achieved by the model with the 3.560761e-05 learning rate, a batch size of 32, dropout value 0.3070714 and attention dropout value 0.2886942758, reaching the validation accuracy of 0.7645, slightly outperforming the baseline BERT model. The validation results of this model are reported in table 3.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
DistilBERT	0.7645	0.7713	0.7645	0.7672

Table 3: Optimized DistilBERT model performance

These results highlight the trade-off between model complexity and computational efficiency. While BERT achieved the highest metrics, DistilBERT, with optimized hyperparameters, demonstrated competitive performance. These findings suggest that lightweight models like DistilBERT can be effectively utilized for political bias detection when computational resources are constrained.

6 Conclusion and Future Work

In this project, we conducted a preliminary analysis and comparison of transformer-based models for political bias detection. Our initial experiments demonstrated that both BERT and DistilBERT deliver comparable performance in this context. However, the significantly lower resource requirements for training DistilBERT make it a more practical choice for extended experimentation and deployment.

The preliminary training results yielded an accuracy of 0.7645, aligning closely with findings from similar studies in this domain. While this performance is promising, there remains substantial room for improvement through more precise fine-tuning of hyperparameters and model configurations.

Additionally, our analysis of n-gram features revealed that the n-grams derived from a larger, domain-specific dataset produced results that differed signiificantly from those of the MBIC dataset. This observation underscores the potential utility of incorporating n-gram features, especially those extracted from diverse and extensive datasets, in future modeling efforts.

Future work could explore several directions to enhance this study. First, employing advanced fine-tuning techniques, such as domain-adaptive pretraining, could significantly improve model performance. Second, experimenting with hybrid models that combine transformer embeddings with n-gram features in a more sophisticated manner may provide deeper insights into the interactions between contextual and statistical features. Finally, expanding the scope of datasets and incorporating more nuanced bias annotations could lead to more robust and generalizable models for political bias detection.

References

[1] Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 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, November 2020. Association for Computational Linguistics.

[2] Wei-Fan Chen, Khalid Al-Khatib, Benno Stein, and Henning Wachsmuth. Detecting media bias in news articles using gaussian bias distributions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4290–4300. Association for Computational Linguistics, 2020.

[3] 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 Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6343–6349, Hong Kong, China, 2019. Association for Computational Linguistics.

[4] Shijia Guo and Kenny Q. Zhu. Modeling multi-level context for informational bias detection by contrastive learning and sentential graph network. *CoRR*, abs/2201.10376, 2022.

[5] Christoph Hube and Besnik Fetahu. Neural based statement classification for biased language. *CoRR*, abs/1811.05740, 2018.

[6] Jan-David Krieger, Timo Spinde, Terry Ruas, Juhi Kulshrestha, and Bela 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 '22, New York, NY, USA, 2022. Association for Computing Machinery.

[7] Yuanyuan Lei and Ruihong Huang. Sentence-level media bias analysis with event relation graph. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 5225–5238, 2024.

[8] 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*, pages 10040–10050, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[9] Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 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, May 2020. European Language Resources Association.

[10] Iffat Maab, Edison Marrese-Taylor, and Yutaka Matsuo. Target-aware contextual political bias detection in news. In *Proceedings of the 13th International Joint Conference on Natural Language Processing (IJCNLP 2023) and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 782–792, Nusa Dua, Bali, November 2023. Association for Computational Linguistics.

[11] Shaina Raza, Deepak John Reji, and Chen Ding. Dbias: Detecting biases and ensuring fairness in news articles. *International Journal of Data Science and Analytics*, 12:1–15, 2022.

[12] Timo Spinde. An interdisciplinary approach for the automated detection and visualization of media bias in news articles. *CoRR*, abs/2112.13352, 2021.

[13] Timo Spinde. An interdisciplinary approach for the automated detection and visualization of media bias in news articles. In *2021 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2021.

[14] Timo Spinde, Smilla Hinterreiter, Fabian Haak, Terry Ruas, Helge Giese, Norman Meuschke, and Bela Gipp. The Media Bias Taxonomy: A Systematic Literature Review on the Forms and Automated Detection of Media Bias. *ACM Computing Surveys [in Review]*, 2023.

[15] Timo Spinde, Jan-David Krieger, Terry Ruas, Jelena Mitrović, Franz Götz-Hahn, Akiko Aizawa, and Bela Gipp. Exploiting transformer-based multitask learning for the detection of media bias in news articles. In Malte Smits, editor, *Information for*

a Better World: Shaping the Global Future, 17th International Conference, iConference 2022, Virtual Event, Proceedings, Part I, volume 13137 of *Lecture Notes in Computer Science*, pages 225–235. Springer, 2022.

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

[17] Timo Spinde, Larissa Rudnitckaia, Kanishka Sinha, Felix Hamborg, Bela Gipp, and Karsten Donnay. Mbic - a media bias annotation dataset including annotator characteristics. *arXiv preprint arXiv:2105.13595*, 2021.

[18] Esther van den Berg and Katja Markert. Context in informational bias detection. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 2020.