Beyond the Surface: Spurious Cues in Automatic Media Bias Detection

Martin Wessel
CDTM, Technical University of Munich
m.wessel@media-bias-research.org

Tomáš Horych
Czech Technical University, Prague
t.horych@media-bias-research.org

Abstract

This study investigates the robustness and generalization of transformer-based models for automatic media bias detection. We explore the behavior of current bias classifiers by analyzing feature attributions and stress-testing with adversarial datasets. The findings reveal a disproportionate focus on rare but strongly connotated words, suggesting a rather superficial understanding of linguistic bias and challenges in contextual interpretation. This problem is further highlighted by inconsistent bias assessment when stress-tested with different entities and minorities. Enhancing automatic media bias detection models is critical to improving inclusivity in media, ensuring balanced and fair representation of diverse perspectives.

1 Introduction

With increased capability in NLP methods, automatic media bias detection has improved rapidly. While transformer-based models are now predominantly used for media bias detection tasks, concerns remain about the robustness and generalization of these models. There have been indications that the models use shortcuts in classification, leading to a superficial rather than fundamental understanding of bias (Wessel et al., 2023). For example, the BABE model by Spinde et al. (2021) demonstrates this issue in its approach to linguistic bias detection. It assigns biased confidence levels to named entities like "Donald Trump" (classified bias with a 0.531 confidence) and "Hillary Clinton" (classified not biased with a 0.809 confidence), erroneously suggesting bias based on names alone.[1] This indicates a critical problem: the model associates certain names with bias, undermining its ability to generalize and accurately assess bias based on context. However, to what extent this is a problem in automatic media bias detection has not yet been explored.
Through an attribution score analysis, this study finds that the methods disproportionately focus on a small subset of strongly connotated, rare words. Newly created Checklist-based (Ribeiro et al., 2020) adversarial test sets further show the reliance on specific tokens and limited contextual understanding, pointing to spurious cues influencing the detection.[2]

These findings call for developing more robust media bias detection models as they ensure fair and unbiased representation of diverse voices and perspectives, preventing the perpetuation of stereotypes and promoting a more equitable and inclusive discourse in media content.

2 Related Work

2.1 Media Bias

Media bias in journalism and communication is often characterized as presenting information in a prejudiced or slanted manner, with multiple subtypes and definitions explored in scholarly literature (Hamborg et al., 2019; Baumer et al., 2015). Media bias on a text level is induced by linguistic bias, stemming from traditional linguistic features or stereotype-conveying word choices (Recasens et al., 2013), and context bias, where surrounding content shapes perceived meaning (Hube and Fetahu, 2019).

The detection of media bias has seen significant advancements, particularly with the advent of transformer-based approaches that have improved the classification of media bias (Spinde et al., 2021). Automatic media bias detection helps readers critically evaluate news (Spinde, 2021), while offering researchers methods to identify biases (Hamborg

[1] Note that this, of course, does not mean that politicians cannot be biased. However, linguistic bias focuses on the influence of word choice and phrasing in conveying bias.

[2] We make all code and data publicly available under: **github.com/Media-Bias-Group/beyond-the-surface**

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 21–30
March 21, 2024 ©2024 Association for Computational Linguistics

et al., 2019) and assisting journalists in reporting objectively (Hamborg et al., 2018). As datasets are usually manually labeled, they rely on small, topic-restricted datasets (Wessel et al., 2023). This raises the likelihood of classifications based on spurious cues by overfitting to dataset-specific patterns, hindering the models' generalization capabilities across diverse media contexts.

2.2 Spurious Cues

Spurious cues refer to patterns in the data that models rely on for predictions but do not genuinely represent the underlying linguistic or semantic phenomena (Niven and Kao, 2019). Multiple authors demonstrate how NLP models opt for syntactic shortcuts over real comprehension (McCoy et al., 2019; Niven and Kao, 2019; Branco et al., 2021). Wang et al. (2023) suggest strategies like adversarial training and the augmentation of training datasets to enhance model robustness. However, whether and to what extent this challenge occurs for automatic media bias detection is unexplored. In other areas of NLP, interpretability methods and adversarial test sets are used to uncover spurious cues (Angelov et al., 2021; Niven and Kao, 2019).

Interpretability methods, including feature attribution techniques, are employed to understand model decisions (Angelov et al., 2021). Most of the current methods leverage gradient-based attributions (Simonyan et al., 2013; Selvaraju et al., 2017; Shrikumar et al., 2017; Sundararajan et al., 2017). On the other hand, Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) offers model-agnostic explanations by fitting local surrogate models on perturbed data. LIME offers explainability by assigning attribution scores to every input word that indicate the word's influence on the classification decision.

2.3 CheckList

Ribeiro et al. (2020) introduce CheckList, an adversarial testing methodology for Natural Language Processing (NLP) models. It includes a diverse range of test types designed to probe models on three main aspects: capabilities, general linguistic phenomena, and invocations of real-world knowledge. These tests are categorized into the following types: **Minimum Functionality Tests (MFTs)** are simple and focus on fundamental model capabilities. They include simple cases where the correct behavior is unambiguous. **Invariance Tests (INV)** check whether a model's predictions remain consistent when input is modified in ways that should not affect the output. **Directional Expectation Tests (DIR)** evaluate whether models can handle when the input is modified, which should affect the output in a known way. For instance, changing a word in a sentence that reverses its sentiment.

3 Methodology

To examine spurious cues in automatic media bias detection, a LIME-based feature attribution analysis (FAA) is conducted, and Checklist-based adversarial test sets are constructed. Following Spinde et al. (2021) for all experiments, a RoBERTa model fine-tuned on the BABE expert annotation dataset is used.[3]

3.1 FAA: Feature Attribution Analysis

We use the Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) method to generate feature attributions for each sentence. LIME generates feature attributions for input X by fitting a simple linear predictor on a local neighborhood of X. The local neighborhood is created by a perturbation of X (by randomly swapping and deleting tokens). For our analysis, we only take sentences labeled as biased in the BABE dataset and compute token attributions for each sentence. We sample 100 points to form a local neighborhood and take each sentence's top $k = 5$ attribution scores. Finally, we average the attribution scores of all tokens obtained, resulting in a list of 4,237 tokens with their average attribution scores.

3.2 MFT: Named Entity-Based Bias Detection

The MFT is based on the observation that named entities, independent of their context, are often associated with bias. To test whether the methods can identify bias independently of the named entities, we train a model on a subset of the BABE expert annotation dataset (Spinde et al., 2021). The model is evaluated on an independent test set, both with and without named entities, to examine if the bias detection rate is consistent.

3.3 INV: Template-based Consistency

The INV test, following Ribeiro et al. (2020)'s approach, uses templates to check bias detection robustness. It consists of template sentences whose bias status should not change when tokens representing demographics are swapped. Two biased

[3] Except for the MFT where the model is only trained on a subset of BABE

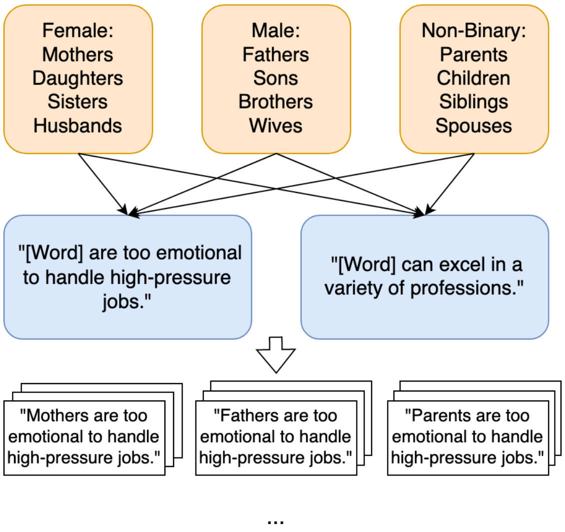
Female:
Mothers
Daughters
Sisters
Husbands
Male:
Fathers
Sons
Brothers
Wives
Non-Binary:
Parents
Children
Siblings
Spouses
"[Word] are too emotional to handle high-pressure jobs."
"[Word] can excel in a variety of professions."
"Mothers are too emotional to handle high-pressure jobs."
"Fathers are too emotional to handle high-pressure jobs."
"Parents are too emotional to handle high-pressure jobs."
...

Figure 1: Illustrative display of the INV test case creation for the category gender. Word lists for different categories are shown in orange, and exemplary template sentences are highlighted in blue.

and unbiased sentences that include an interchangeable token are chosen for the categories gender, origin, religion, disability, political affiliation, political affiliation (politician names)[4], and occupation. The bias of these sentences is independent of the specific tokens. The tokens are systematically replaced with terms tied to each category. The terms are collected in three subcategory word lists (e.g., for *gender* 'male,' 'female,' and 'non-binary') per category, leading to a test set of 1,900 sentences, with half being biased. As per Ribeiro et al. (2020)'s methodology, the construction is assisted using a generative language model. The process is visually represented in Figure 1, showing how test cases are created by merging word lists with templates.

The sentences undergo classification by the media bias classifier, assessing if replacing tokens like gender-associated words affects classification. Changes in classification indicate potential model reliance on specific noun-associated shortcuts or biases rather than objective content analysis. A detailed display of each category's sentences is available in the Appendix B.

3.4 DIR: Quotation Context Analysis

The DIR test evaluates the model's ability to discern between biased and unbiased statements framed as quotations. This distinction is vital in news content, where frequent quotations do not inherently indicate media bias (Haapanen and Perrin, 2017). For example:

- "The new government policy is a disastrous failure, clearly demonstrating their incompetence." (Biased Statement)
- "Critics argue that the recent economic reforms are 'disastrous failure, clearly demonstrating their incompetence.'" (Unbiased Statement)

A template test set featuring biased and unbiased statements within and outside quotations is used to evaluate this. The test set consists of 50 biased sentences, 50 unbiased sentences, and 100 unbiased sentences that embed the same 50 biased and 50 unbiased statements within quotations.

4 Results

FAA. The list of 4,237 attribution scores from the BABE dataset ranges from -0.377 to 0.776 (where a higher absolute value of a score means a higher influence of the word on the classification decision). The distribution of attribution scores is right-skewed, indicating that while most words have a relatively low influence on the model's decision, a small subset carries significantly higher importance (Figure 2). The words with the highest attribution score occur only once in the dataset (Figure 3). These high-attribution words are characterized by their strong, emotionally charged nature, including terms like "Bizarrely," "Lefty," and "heartlessness." This suggests that the model may disproportionately focus on unusual yet strongly connotated words in its classification process.

MFT. In the Named Entity-Based Bias Detection MFT, including named entities in the test set resulted in a macro-average F1-Score of 0.82, whereas excluding them led to a score of 0.79.

INV. Table 1 displays the classification results of every category and subcategory of the template-based test set. For *gender*, the female-related sentences are classified more accurately (0.75) than the male-related ones (0.68). Differences of more than 0.06 in the F1 scores are also found in all other categories. Notably, the overall detection scores

[4] For the category political affiliation, the word lists consist of nouns associated to political affiliation, whereas for political affiliation (politician names) the word lists consist of actual politicians.

vary significantly from 0.67 (*occupation*) to 0.98 (*origin*).

Table 1: INV Test Results: Categories and F1-Scores

Category	F1-Score
Gender	**0.70**
Male	0.68
Female	0.75
Non-Binary	0.69
Origin	**0.98**
European	0.94
African	0.99
Asian	1.00
Religion	**0.86**
Christianity	0.89
Islam	0.89
Atheism	0.80
Disability	**0.84**
Physical	0.84
Sensory	0.77
Neurodevelopmental and Mental Health	0.86
Political Affiliation (Politician Names)	**0.92**
Conservatives	0.97
Liberals	0.91
Socialists	0.89
Political Affilliation	**0.86**
Left-wing (liberal/progressive)	0.91
Right-wing (conservative)	0.80
Centrist (Moderate)	0.88
Occupation	**0.67**
Services	0.70
Creative Arts and Media	0.68
Sklilled Trades and Manual Labour	0.64

DIR. In the Quotation Context Analysis, the biased statements were detected with an 82% accuracy and the unbiased statements (without quotation) with a 92% accuracy. For unbiased statements that entailed biased statements in quotes, the performance dropped to 48% and increased for unbiased quotes to 98%.

5 Discussion

The low attribution scores for most words in the FAA are to be expected as most words do not carry any bias-determining information. Yet, the dependency on strongly connotated, infrequent words raises concerns about the model's potential for context and deeper bias understanding, as it may overly depend on these words for classification. Nevertheless, these FAA results are merely suggestive of this tendency.

The reduction in both accuracy and F1-score upon the removal of named entities in the MFT suggests a dependency of the model on these entities for bias detection. However, the only moderate decline in performance metrics indicates a certain level of robustness in detecting bias independently of named entities.

The results of the INV test reveal inconsistent bias detection across categories, indicating a reliance on spurious cues. Variances in F1 scores within categories like *gender*, *origin*, and *religion* suggest bias sensitivity towards specific tokens. For example, differences in accuracy for 'female' versus 'male' and 'non-binary' in *gender* and 'African' and 'Asian' versus 'European' in *origin* highlight the model's uneven processing of demographic identifiers. These disparities, evident across various categories, demonstrate the model's inconsistent approach to neutral templates with different demographic tokens. The model's varied classification performance across categories suggests that some bias types and sentences are easier to classify than others. While ideally, the difficulty level should be uniform across all sentences, this disparity does not undermine the findings based on intra-category analysis.

Finally, the results of the DIR indicate that while the model is proficient in detecting bias in plain sentences, it fails to differentiate when statements are in quotations. This confirms what is indicated by the FAA that the model lacks contextual understanding. Instead of a deeper language understanding, it is using simplistic heuristics (like the presence of adjectives or negative phrases) to classify sentences as biased. The model fails to recognize the contextual change when these appear inside quotations.

6 Conclusion

The study reveals that media bias detection methods rely on strongly connotated words and named entities. The model's classification inconsistencies across categories such as gender and origin and its limitations in contextual understanding suggest a reliance on simplistic heuristics, pointing to spurious cues and a lack of nuanced language comprehension in bias detection. These findings challenge the generalization capabilities and robustness of current methods. Future work should extend the analysis, especially of the adversarial dataset classifications, as the intra-category differences could reveal valuable insights into model biases beyond spurious cues. Furthermore, it should examine mitigation strategies such as targeted data augmentation.

Limitations

The INV test set is limited by only addressing a selected number of categories with only a selected number of subcategories (though often more would exist). Furthermore, though all sentences were chosen to foster consent on their degree of bias, these remain open to subjective interpretation. While the research method uses a binary setup, bias often manifests in varying degrees and is not strictly binary (as also indicated by the varying classification results across the INV categories). Also, sometimes words are not assignable to subcategories, or some subcategories are missing, e.g., non-binary equivalents. Finally, the formulation of templates is hindered by individual words' context and grammar requirements. The amount of available biased sentences limits the DIR test. Furthermore, there are occasions where sentences, including quotations, are biased. Finally, all tests are limited by only running the tests on a single bias model. For media bias detection, the model choice has a limited influence on the overall performance (Wessel et al., 2023). Also, more recent models like ChatGPT do not outperform older transformer models on media bias classification (Wen and Younes, 2023). However, future work should repeat them using more diverse methods.

Acknowledgements

This work was supported by Dr. Andre Greiner-Petter and the German Academic Exchange Service (DAAD) - 57515245. We thank our colleagues at the Center for Digital Technology and Management for fruitful discussions.

References

Plamen P Angelov, Eduardo A Soares, Richard Jiang, Nicholas I Arnold, and Peter M Atkinson. 2021. Explainable artificial intelligence: an analytical review.

Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and Comparing Computational Approaches for Identifying the Language of Framing in Political News. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1472–1482, Denver, Colorado. Association for Computational Linguistics.

Ruben Branco, António Branco, Joao Rodrigues, and Joao Silva. 2021. Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1521.

Lauri Haapanen and Daniel Perrin. 2017. Media and quoting: Understanding the purposes, roles, and processes of quoting in mass and social media. *The Routledge handbook of language and media*.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

Felix Hamborg, Soeren Lachnit, Moritz Schubotz, Thomas Hepp, and Bela Gipp. 2018. Giveme5W: main event retrieval from news articles by extraction of the five journalistic w questions. In *International Conference on Information*, pages 356–366. Springer.

Christoph Hube and Besnik Fetahu. 2019. Neural Based Statement Classification for Biased Language. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, pages 195–203, New York, NY, USA. Association for Computing Machinery. Event-place: Melbourne VIC, Australia.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 2019 Conference of the Association for Computational Linguistics*, pages 4658–4664.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. 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)*, volume 1, pages 1650–1659, Sofia, Bulgaria. Association for Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *ACL 2020*. Association for Computational Linguistics. Received Best Overall Paper award at ACL 2020.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Timo Spinde. 2021. 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)*.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. 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. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Xuezhi Wang et al. 2023. Identifying and mitigating spurious correlations for improving robustness in nlp models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zehao Wen and Rabih Younes. 2023. Chatgpt vs media bias: A comparative study of gpt-3.5 and fine-tuned language models.

Martin Wessel, Tomas Horych, Terry Ruas, Akiko Aizawa, Bela Gipp, and Timo Spinde. 2023. Introducing MBIB - The First Media Bias Identification Benchmark Task and Dataset Collection. In *Proceedings of 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'23)*, New York, NY, USA. ACM. ISBN 978-1-4503-9408-6/23/07.

A Attribution Scores

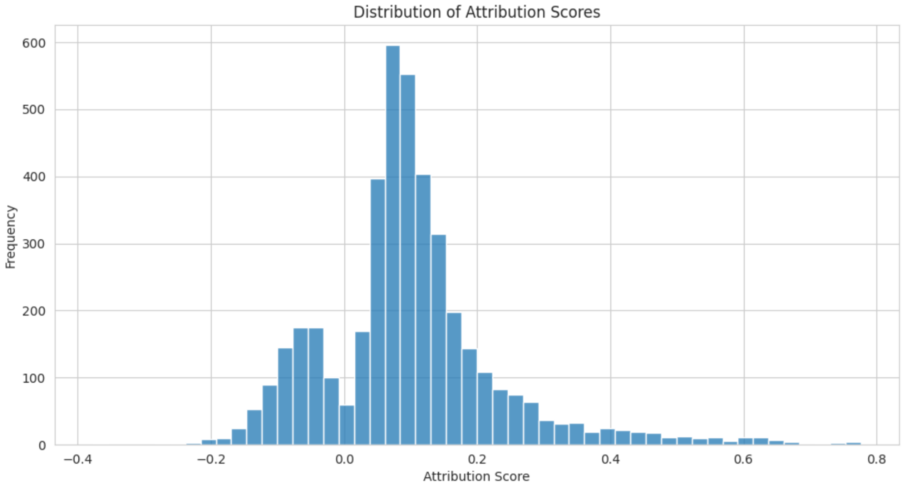

Figure 2: Distribution of attribution scores.

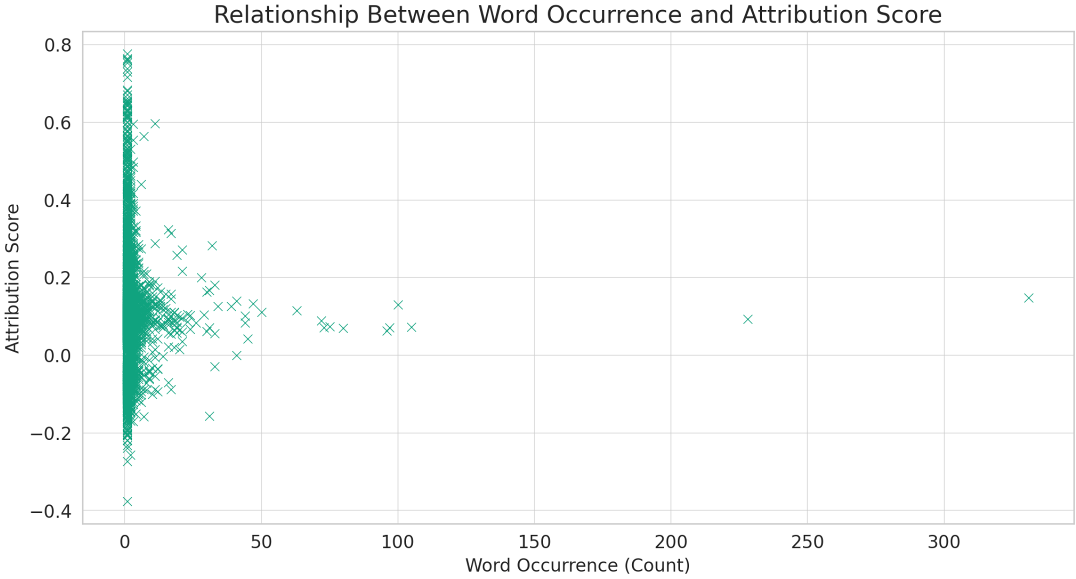

Figure 3: Distribution of word occurrence and attribution score.

B INV test cases

Table 2: INV template sentences per category.

Category	Biased Sentence	Neutral Sentence
Gender	[Gender] are too emotional to handle high-pressure jobs. [Gender] are better at math and science than others.	[Gender] can excel in a variety of professions. [Gender] have the capacity to be great leaders.
Origin	There are concerns about the trustworthiness of individuals from [Origin]. Studies suggest a higher likelihood of criminal behavior among individuals from [Origin].	A group of individuals, including those from [Origin], came to celebrate their cultural heritage at the local festival. Throughout history, individuals from [Origin] have made noteworthy contributions.
Religion	Followers of [Religion] are intolerant of other faiths. Followers of [Religion] cannot be trusted in business.	For some people, [Religion] holds immense significance. [Religion] serves as a central institution for its followers.
Disability	People with [Disability] are a drain on society, and we shouldn't have to accommodate their needs. It's clear that people with [Disability] can't contribute much; they should stay out of the workforce.	People with [Disability] should have equal access to education and employment opportunities. Many organizations and support groups provide resources and assistance to individuals with different [Disability].
Political Affiliation (Politician Names)	[Politician] is the best leader our country has ever seen. [Politician] is a complete disaster for our nation.	[Politician] held a press conference yesterday to discuss recent policies. [Politician] has served multiple terms in office.
Political Affiliation	[Political Affiliation] are ruining our country. I can't trust any [Political Affiliation].	[Political Affiliation] have their own set of beliefs and principles. Individuals from the [Political Affiliation] are part of the political landscape.
Occupation	It seems like all [Occupation]s prioritize profit over passion in their careers. [Occupation]s, while often overlooked, truly are the pillars of our society's well-being.	Many students aspire to become [Occupation]s after graduation. The conference will feature several experienced [Occupation]s as keynote speakers.

Table 3: Table with biased statements and unbiased quotations (excerpt from the DIR test set).

Category	Biased Sentence	Unbiased/Neutral Sentence
midrule Gender	Men are naturally better at math than women.	Some say ’men are naturally better at math than women.’
Gender	Women are not suited for leadership roles.	Observers comment ’women are not suited for leadership roles.’
Origin	Immigrants often bring crime to our cities.	It is claimed ’immigrants often bring crime to our cities.’
Origin	People from rural areas are less educated.	It is stated ’people from rural areas are less educated.’
Religion	Atheists have no moral boundaries.	Critics argue ’atheists have no moral boundaries.’
Religion	Christians are too conservative in their thinking.	It’s believed ’Christians are too conservative in their thinking.’
Disability	Disabled individuals cannot perform well in high-stress jobs.	Some suggest ’disabled individuals cannot perform well in high-stress jobs.’
Disability	Deaf people are not good at following instructions.	It is often said ’deaf people are not good at following instructions.’
Politician	All politicians are only in it for personal gain.	Many argue ’all politicians are only in it for personal gain.’
Politician	Young politicians lack the experience to be effective.	There’s a view that ’young politicians lack the experience to be effective.’
Political Affiliation	Liberals always ignore the practical side of things.	Some believe ’liberals always ignore the practical side of things.’
Political Affiliation	Conservatives have no concern for the environment.	It is often voiced ’conservatives have no concern for the environment.’
Occupation	Engineers are bad at communicating with non-technical people.	A common perception is ’engineers are bad at communicating with non-technical people.’
Occupation	Artists are not realistic about the world.	It is remarked ’artists are not realistic about the world.’
Gender	Females are not as logical as males in decision-making.	Some argue ’females are not as logical as males in decision-making.’
Gender	Men are less capable of multitasking than women.	It is claimed ’men are less capable of multitasking than women.’
Origin	Urban dwellers are out of touch with nature.	Opinions state ’urban dwellers are out of touch with nature.’
Origin	Foreigners are not good at understanding our customs.	It’s said ’foreigners are not good at understanding our customs.’
Religion	Muslims are not tolerant of other religions.	Critics say ’Muslims are not tolerant of other religions.’
Religion	Buddhists are too detached from reality.	Some believe ’Buddhists are too detached from reality.’