A Taxonomy of Mitigation Methods for Hate Speech Detection

Jan Fillies
Freie Universität Berlin
Berlin, Germany
Institut für Angewandte Informatik
Leipzig, Germany
fillies@infai.org

Marius Wawerek
Freie Universität Berlin
Berlin, Germany

Adrian Paschke
Freie Universität Berlin
Berlin, Germany
Institut für Angewandte Informatik
Leipzig, Germany
Fraunhofer FOKUS
Berlin, Germany

Abstract

This document presents more accessible versions of the bias mitigation taxonomy presented in the original publication “A Comprehensive Taxonomy of Bias Mitigation Methods for Hate Speech Detection”, submitted at WOAH 2025.

This document presents more accessible versions of the bias mitigation taxonomy presented in the original publication “A Comprehensive Taxonomy of Bias Mitigation Methods for Hate Speech Detection”, submitted at WOAH 2025.

There are a total of 4 figures presented here. Figure 1 is the original (short) taxonomy presented in the main text of the publication.

Figure 2 is the extended version of the taxonomy, with example methods per class and citations, presented in the appendix of the publication.

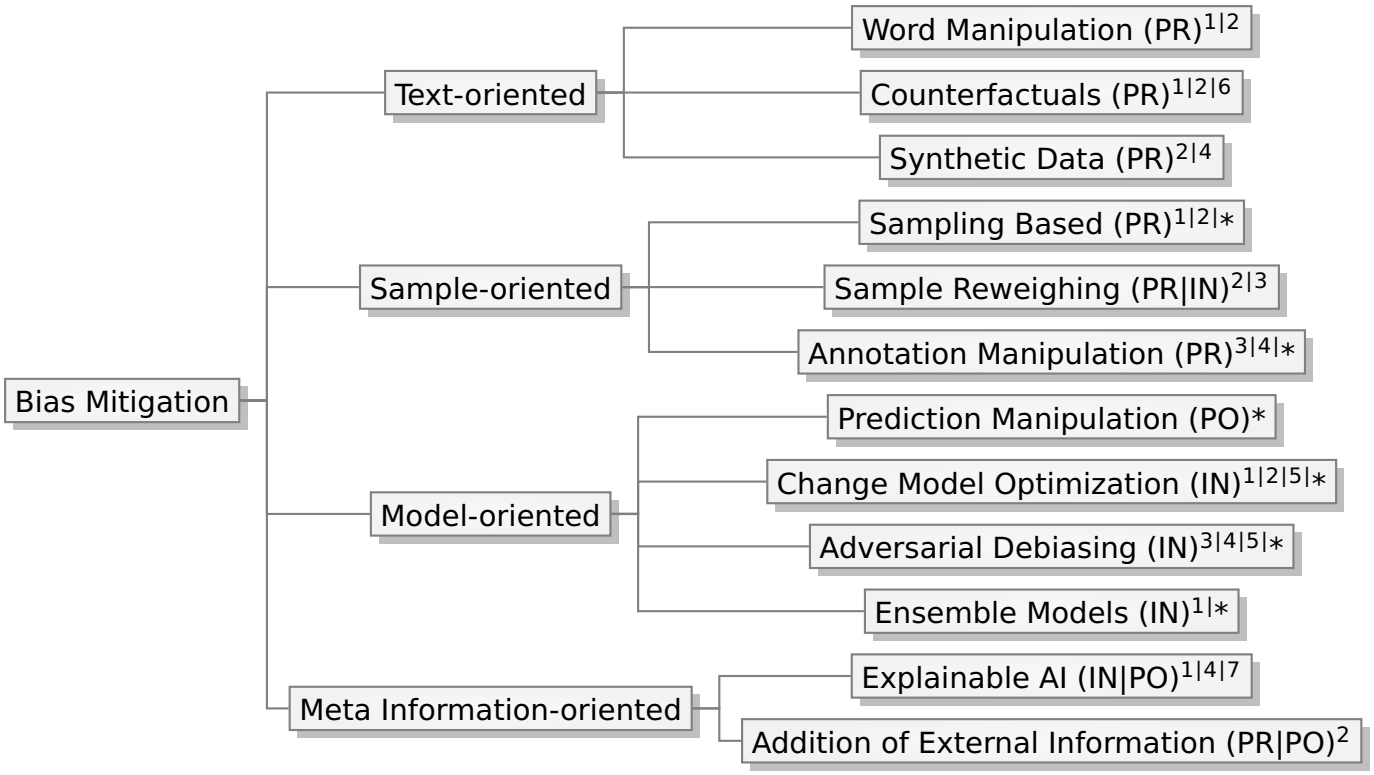

Figure 1: Taxonomy of Bias Mitigation Methods based on their Principle of Operation. Each symbol marks a class of processing PR = Pre-processing, IN = In-Processing, PO = Post-Processing, Historical Bias = [1], Representation Bias = [2], Measurement Bias = [3], Aggregation Bias = [4], Learning Bias = [5], Evaluation Bias = [6], Deployment Bias = [7], Requires knowledge about the protected attribute = *

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. *arXiv preprint*.

[2] Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust Hate Speech Detection in Social Media: A Cross-Dataset Empirical Evaluation.

[3] Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based Attention Regularization Frees Unintended Bias Mitigation from Lists.

[4] Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical Bias Removal for Hate Speech Detection Task using Knowledge-based Generalizations. In *The World Wide Web Conference*, pages 49–59, New York, NY, USA. ACM.

[5] Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2021. Differential Tweetment: Mitigating Racial Dialect Bias in Harmful Tweet Detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 116–128, New York, NY, USA. ACM.

[6] Yi Cai, Arthur Zimek, Gerhard Wunder, and Eirini Ntoutsi. 2022. Power of Explanations: Towards automatic debiasing in hate speech detection.

[7] Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Mohammad Atari, Xiang Ren, and Morteza Dehghani. 2020. Fair Hate Speech Detection through Evaluation of Social Group Counterfactuals.

[8] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, New York, NY, USA. ACM.

[9] Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech.

[10] Yacine Gaci, Boualem Benatallah, Fabio Casati, and Khalid Benabdeslem. 2022.

Debiasing Pretrained Text Encoders by Paying Attention to Paying Attention. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9582–9602, Stroudsburg, PA, USA. Association for Computational Linguistics.

[11] Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2022. Handling Bias in Toxic Speech Detection: A Survey.

[12] Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Diverse Adversaries for Mitigating Bias in Training.

[13] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection.

[14] Przemyslaw Joniak and Akiko Aizawa. 2022. Gender Biases and Where to Find Them: Exploring Gender Bias in Pre-Trained Transformer-based Language Models Using Movement Pruning.

[15] Ratnesh Kumar Joshi, Arindam Chatterjee, and Asif Ekbal. 2023. Saliency Guided Debiasing: Detecting and mitigating biases in LMs using feature attribution. *Neurocomputing*, page 126851.

[16] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33.

[17] Faisal Kamiran, Sameen Mansha, Asim Karim, and Xiangliang Zhang. 2018. Exploiting reject option in classification for social discrimination control. *Information Sciences*, 425:18–33.

[18] Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. *Contextualizing Hate Speech Classifiers with Post-hoc Explanation*. Association for Computational Linguistics.

[19] Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. *Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application*. arXiv.

[20] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-aware Classification. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, pages 853–862, New York, New York, USA. ACM Press.

[21] Ulrike Kuhl, André Artelt, and Barbara Hammer. 2023. *For Better or Worse: The Impact of Counterfactual Explanations' Directionality on User Behavior in xAI*, volume 1903. Springer, Cham.

[22] Yunyi Li, Maria De-Arteaga, and Maytal Saar-Tsechansky. 2023. Mitigating Label Bias via Decoupled Confident Learning.

[23] Pranay K. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, and Ruchir Puri. 2019. Bias Mitigation Post-processing for Individual and Group Fairness. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2847–2851. IEEE.

[24] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

[25] Francimaria R.S. Nascimento, George D.C. Cavalcanti, and Márjory Da Costa-Abreu. 2022. Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on social media using ensemble learning. *Expert Systems with Applications*, 201:117032.

[26] Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. 2020. SSMBA: Self-Supervised Manifold Based Data Augmentation for Improving Out-of-Domain Robustness. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1268–1283.

[27] Nicolas Ocampo, Elena Cabrio, and Serena Villata. 2023. Playing the Part of the Sharp Bully: Generating Adversarial Examples for Implicit Hate Speech Detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2758–2772, Toronto, Canada. Association for Computational Linguistics.

[28] Ebuka Okpala, Long Cheng, Nicodemus Mbwambo, and Feng Luo. 2022. AAEBERT: Debiasing BERT-based Hate Speech Detection Models via Adversarial Learning. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE.

[29] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. *Reducing Gender Bias in Abusive Language Detection*. Association for Computational Linguistics.

[30] Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. 2019. Detecting and Monitoring Hate Speech in Twitter. *Sensors (Basel, Switzerland)*, 19(21).

[31] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. *arXiv preprint*.

[32] Muhammad Deedahwar Mazhar Qureshi, M. Atif Qureshi, and Wael Rashwan. 2023. Toward Inclusive Online Environments: Counterfactual-Inspired XAI for Detecting and Interpreting Hateful and Offensive Tweets. In *Explainable Artificial Intelligence*, Communications in Computer and Information Science, pages 97–119, Cham. Springer Nature Switzerland and Imprint Springer.

[33] Alan Ramponi and Sara Tonelli. 2022. Features or Spurious Artifacts? Data-centric Baselines for Fair and Robust Hate Speech Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3027–3040. Association for Computational Linguistics.

[34] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. *arXiv preprint*.

[35] Paras Sheth, Tharindu Kumarage, Raha Moraffah, Aman Chadha, and Huan Liu. 2023. Causality Guided Disentanglement for Cross-Platform Hate Speech Detection.

[36] Paras Sheth, Tharindu Kumarage, Raha Moraffah, Aman Chadha, and Huan Liu. 2023. PEACE: Cross-Platform Hate Speech Detection- A Causality-guided Framework.

[37] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting Racial Bias in Hate Speech Detection.

[38] Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative Data Augmentation for Commonsense Reasoning. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025.

[39] Xuan Zhao, Simone Fabbrizzi, Paula Reyero Lobo, Siamak Ghodsi, Klaus Broelemann, Steffen Staab, and Gjergji Kasneci. 2023. *Adversarial Reweighting Guided by Wasserstein Distance for Bias Mitigation*. arXiv.

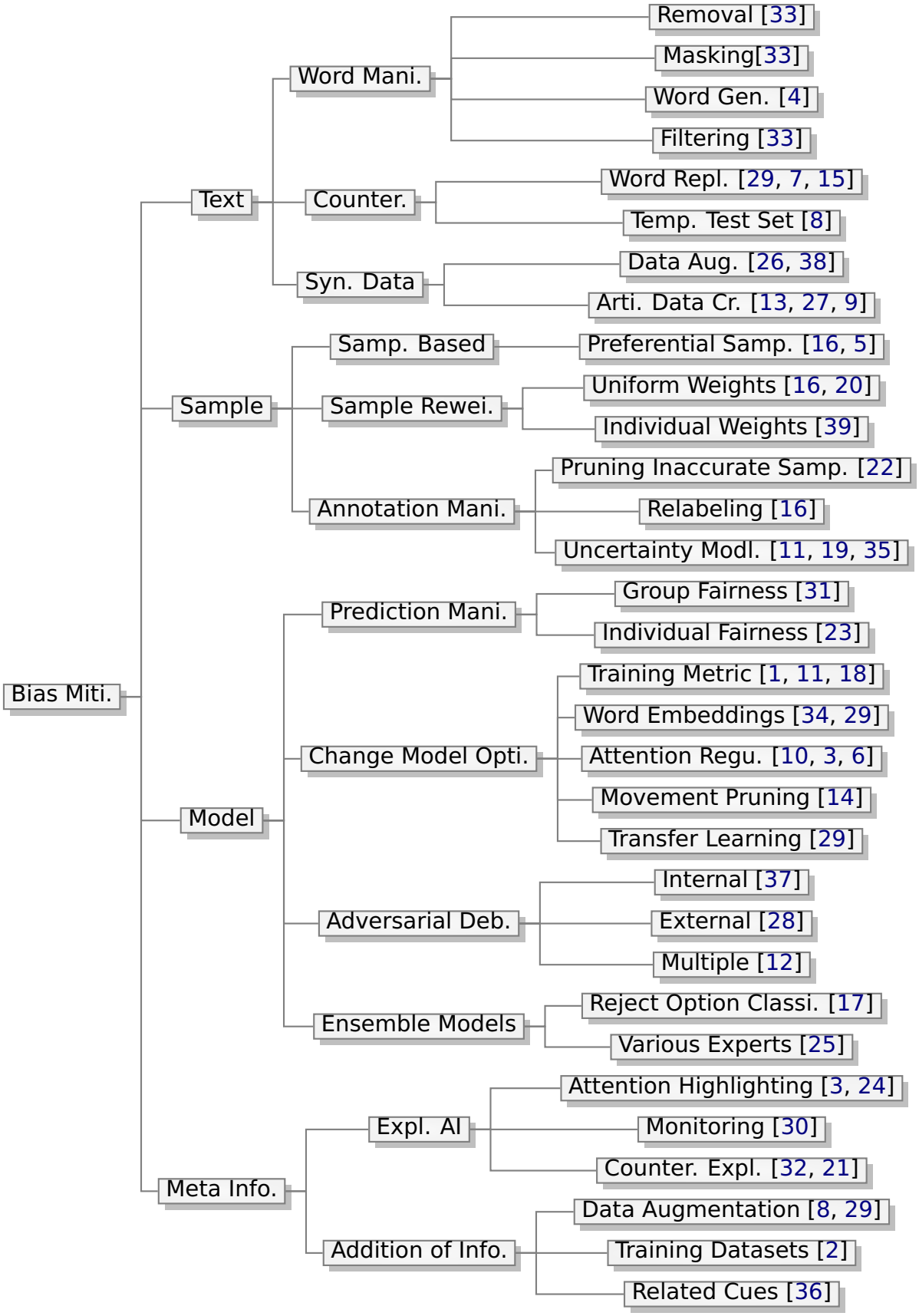

Figure 2: Class broken down in concrete approaches with citations. Miti. = Mitigation, Mani. = Manipulation, Gen. = Generalization, Counter. = Counterfactuals, Repl. = Replacements, Temp. = Template, Syn. = Synthetic, Aug. = Augmentation, Arti. = Artificial, Cr. = Creation, Samp. = Sampling, Rewei. = Reweighing, Opti. = Optimization, Deb = Debiasing, Model. = Modeling, Expl. = Explainable, Regu. = Regularization, Classi. = Classifaction