GENDER BIAS DETECTION IN NATURAL LANGUAGE USING EXPLAINABLE BERT–BILSTM–ATTENTION AND ROBERTA–CNN–BIGRU HYBRID MODELS

Akash Saraswat[1], Dr. Dinesh Chandra Misra[2], Dr. Rahul Kumar[3]

PhD Scholar[1], Associate Professor[2,3]

[1,2]Department of Computer Science and Engineering, Dr. K. N. Modi University
[3]Department of Computing and Informatics, Sir Padampat Singhania University, Udaipur Rajasthan
sharma9910483702@gmail.com[1], dcmishra99@gmail.com[2] , rahulkumar1680@gmail.com[3]

Abstract

Gender bias in natural language presents a significant challenge for the development of fair, reliable, and socially responsible Natural Language Processing (NLP) and Large Language Model (LLM) systems. Gender stereotypes, gender-marked expressions, and implicitly biased language can be embedded within textual data and subsequently reproduced by automated language technologies. Therefore, accurate and interpretable detection of gender bias is essential for evaluating the fairness of language models and supporting the development of effective bias-mitigation strategies. This research proposes an explainable hybrid deep learning framework for automated gender-bias detection in natural language. The study investigates gender bias using appropriate gender-bias datasets containing sentences representing gender-marked, stereotypical, neutral, and potentially biased linguistic patterns. Two complementary hybrid architectures are developed and evaluated. The first model combines Bidirectional Encoder Representations from Transformers (BERT), Bidirectional Long Short-Term Memory (BiLSTM), and an Attention mechanism to capture contextual representations, sequential dependencies, and the most influential linguistic features contributing to classification decisions. The second model integrates Robustly Optimized BERT Pretraining Approach (RoBERTa), Convolutional Neural Network (CNN), and Bidirectional Gated Recurrent Unit (BiGRU) to capture contextual, local, and sequential patterns associated with gender-biased language.

The proposed models are systematically evaluated and compared using accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix, and other relevant evaluation measures. Particular emphasis is placed on explainability so that the system can provide insights into the linguistic patterns influencing its predictions rather than functioning solely as a black-box classifier. In addition, an interactive detection system is developed to enable users to enter unseen sentences and obtain automated gender-bias predictions. The comparative analysis is intended to identify the strengths and limitations of the proposed architectures and determine their suitability for practical fairness assessment. The research provides an integrated framework combining transformer-based contextual learning, sequential deep learning, attention-based interpretability, and interactive prediction for gender-bias detection. The resulting framework can provide a practical foundation for auditing NLP and LLM evaluation data and can support future research focused on explainable artificial intelligence, responsible AI, and automated gender-bias mitigation.

Keywords: Gender Bias Detection, Natural Language Processing, Large Language Models, BERT, BiLSTM, Attention Mechanism, RoBERTa, CNN, BiGRU, Explainable AI, Hybrid Deep Learning, Transformer Models, Responsible AI, Bias Mitigation

1. Introduction

Natural Language Processing (NLP) has become a fundamental component of modern artificial intelligence (AI), enabling machines to analyse, understand, generate, and interact with human language. The rapid development of transformer-based architectures and Large Language Models (LLMs) has substantially improved the performance of NLP systems across applications such as machine translation, sentiment analysis, information retrieval, question answering, text generation, and conversational AI. Models such as BERT and RoBERTa have demonstrated strong capabilities in learning contextual representations from large-scale textual corpora, while subsequent LLMs have extended these capabilities to increasingly complex language tasks (Devlin et al., 2019; Liu et al., 2019; Zhao et al., 2023). However, the effectiveness of these systems is closely related to the data on which they are trained and evaluated. Since language data reflects social attitudes, cultural norms, historical inequalities, and stereotypes, AI systems can learn and reproduce undesirable social biases, including gender bias.

Gender bias in language refers to linguistic patterns that associate particular characteristics, behaviours, occupations, social roles, or expectations disproportionately with a particular gender. Such bias may appear explicitly through discriminatory statements or implicitly through stereotypes and unequal associations. For example, descriptions of women may be disproportionately associated with appearance, family, or emotional characteristics, whereas descriptions of men may be associated more frequently with leadership, technical ability, or professional competence. When these patterns occur in training and evaluation datasets, computational models may learn statistical associations that reinforce existing stereotypes. Research has demonstrated that pretrained language models can encode social and demographic biases present in their underlying corpora, creating concerns regarding fairness and responsible deployment (Caliskan, Bryson and Narayanan, 2017; Nangia et al., 2020; Blodgett et al., 2020).

The problem becomes particularly important with the increasing adoption of LLMs in education, recruitment, healthcare, customer service, content generation, and decision-support applications. Bias embedded within language models can influence generated responses, classifications, recommendations, and evaluations, potentially producing systematically different outcomes for texts referring to different genders. Recent research has consequently emphasised the importance of evaluating fairness not only in model architecture but also in datasets, benchmarks, prompts, and generated outputs (Gallegos et al., 2024; Navigli et al., 2023). Bias evaluation is challenging because gender bias is not always represented through obvious discriminatory vocabulary. It can be contextual, subtle, culturally dependent, and expressed through relationships between words and broader linguistic structures. Therefore, conventional keyword-based approaches may be insufficient for detecting complex forms of gender bias.

Deep learning and transformer-based NLP techniques provide promising approaches for addressing this challenge. BERT generates bidirectional contextual representations that allow words to be interpreted according to their surrounding context (Devlin et al., 2019). RoBERTa further improves transformer pretraining through modifications to the training procedure and has achieved strong performance across numerous language understanding tasks (Liu et al., 2019). Nevertheless, transformer representations can be complemented by recurrent and convolutional architectures. BiLSTM and BiGRU networks can capture

sequential dependencies in both directions, while CNNs can identify informative local linguistic patterns. Attention mechanisms can additionally assign greater importance to relevant parts of an input sentence, supporting both classification performance and interpretability.

Explainability is particularly important when detecting socially sensitive bias. A model that simply labels a sentence as biased may provide limited value if users cannot understand which linguistic patterns contributed to the decision. Explainable AI (XAI) techniques can therefore improve transparency, facilitate model auditing, and assist researchers in identifying problematic linguistic features. This research addresses these challenges by developing and comparing two hybrid architectures: an explainable BERT–BiLSTM–Attention model and a RoBERTa–CNN–BiGRU model. The proposed framework aims to combine contextual transformer representations with sequential, local-feature, and attention-based learning for reliable gender-bias detection. The study further evaluates the models using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices, while an interactive system is proposed for testing unseen sentences. Through this integrated approach, the research seeks to contribute to automated fairness assessment in NLP and LLM evaluation data and establish a practical foundation for future research on responsible and bias-aware artificial intelligence (Weidinger et al., 2022; Gallegos et al., 2024).

Objectives of the Research

1. To identify and analyse gender bias in natural language and LLM evaluation datasets.
2. To develop a BERT–BiLSTM–Attention model for detecting gender-biased language.
3. To develop a RoBERTa–CNN–BiGRU model for automatic gender-bias detection.
4. To compare the performance of both proposed models using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix.
5. To develop an interactive system that can detect gender bias in new or unseen sentences.

2. Literature Review

Table 2.1: Comparative Literature Review of Recent Studies on Gender and Textual Bias

Author Name (Year)	Main Concept	Area	Model Used	Advantages	Limitations
Donald et al. (2026)	Automated textual bias detection and generalisation	Textual Bias	XLNet, DistilBERT, RoBERTa, ELECTRA	Compares multiple Transformers and real-world data	Performance decreases on real-world data; not gender-specific
Bryskina et al. (2026)	Detection of multiple textual biases	Gender, Race, Religion	11 lightweight LLMs	Large dataset with 8,745 sentences	Zero-shot accuracy remained below 70%; implicit bias is difficult
Li, Yan and Dong (2026)	Context-based hierarchical bias detection	Media Bias	RoBERTa + Transformer Aggregator	Uses document context and multi-task classification	Computationally complex; not gender-specific
Park et al. (2026)	Gender-sensitivity evaluation benchmark	Gender / Political Discourse	GPT-4.1, GPT-4o	Korean benchmark with 6,024 utterances	Limited to Korean political discourse
Salas-Jimenez et al. (2026)	LLM-based bias detection and mitigation	General / Multilingual Bias	Llama, Gemma, SBERT + SVM	Combines LLMs with traditional classifiers	Complex methodology and prompt dependence
Mohammadi et al. (2026)	Reliability-aware sexism detection	Multilingual Sexism	RA-DPO, GPT-4o, Llama, Qwen	Uses uncertainty and annotator agreement	Requires annotation-agreement data; computationally demanding
Sabir,	Confidence	Gender Bias	Six LLMs	Introduces	Focused mainly

Kängsepp and Sharma (2026)	and gender-bias analysis			Gender-ECE for calibration	on pronoun resolution
Liu and Li (2026)	Gender-stereotype diagnosis and mitigation	Japanese Gender Bias	Japanese BERT-based PLMs	Provides gender-bias evaluation and mitigation measures	Limited to Japanese NLP and specific tasks

The literature reviewed in Table 2.1 shows that recent research has increasingly focused on detecting, evaluating, and mitigating bias in natural language and Large Language Models. Donald et al. (2026) and Li, Yan and Dong (2026) demonstrate the effectiveness of Transformer-based models for general textual and media bias detection, but their approaches are not specifically designed for gender bias. Bryskina, Songailaitė and Mandravickaitė (2026) provide broader gender, race, religion, and appearance bias evaluation, although the relatively low zero-shot accuracy indicates the difficulty of identifying implicit and contextual bias. Park et al. (2026) and Liu and Li (2026) focus more specifically on gender sensitivity and gender stereotypes, but their studies are restricted to particular linguistic or cultural settings. Similarly, Sabir, Kängsepp and Sharma (2026) investigate confidence calibration rather than developing a comprehensive gender-bias classifier. Mohammadi et al. (2026) introduce uncertainty and annotator agreement into sexism detection, providing a more reliability-oriented approach, while Salas-Jimenez et al. (2026) combine LLM representations with conventional machine-learning classifiers and mitigation techniques. Overall, the reviewed studies indicate that Transformer and LLM-based approaches provide strong capabilities for bias analysis, but important gaps remain in general-purpose gender-bias classification, explainability, hybrid deep-learning architectures, and practical detection of unseen sentences. These limitations motivate the present research, which proposes an explainable BERT–BiLSTM–Attention model and a RoBERTa–CNN–BiGRU hybrid model and compares their performance using multiple evaluation metrics.

3. Research Methodology

This research adopts a systematic experimental methodology to develop, evaluate, and compare deep learning models for automated gender-bias detection in natural language. The overall methodology consists of dataset selection, data preprocessing, exploratory analysis, dataset preparation, transformer-based feature extraction, hybrid model development, model training, performance evaluation, explainability analysis, and interactive system development. The primary objective of this methodology is to identify gender-marked or potentially biased language accurately while also examining the ability of the proposed models to generalise to unseen sentences. The research follows a comparative experimental design in which two hybrid architectures, BERT–BiLSTM–Attention and RoBERTa–CNN–BiGRU, are developed under comparable experimental conditions.

The first stage involves collecting appropriate datasets containing textual examples related to gender bias, stereotypes, sexism, and neutral language. The selected datasets are examined for class labels, duplicate records, missing values, class distribution, sentence length, and linguistic characteristics. Where multiple datasets are available, they can be combined after ensuring that their label definitions are compatible. The data is divided into training, validation, and testing sets, typically using a stratified division to maintain similar class proportions across the subsets. This prevents information leakage and ensures that the final test set contains previously unseen examples for objective evaluation. Exploratory analysis is also performed to understand the frequency of biased and non-biased sentences and to identify potential class imbalance.

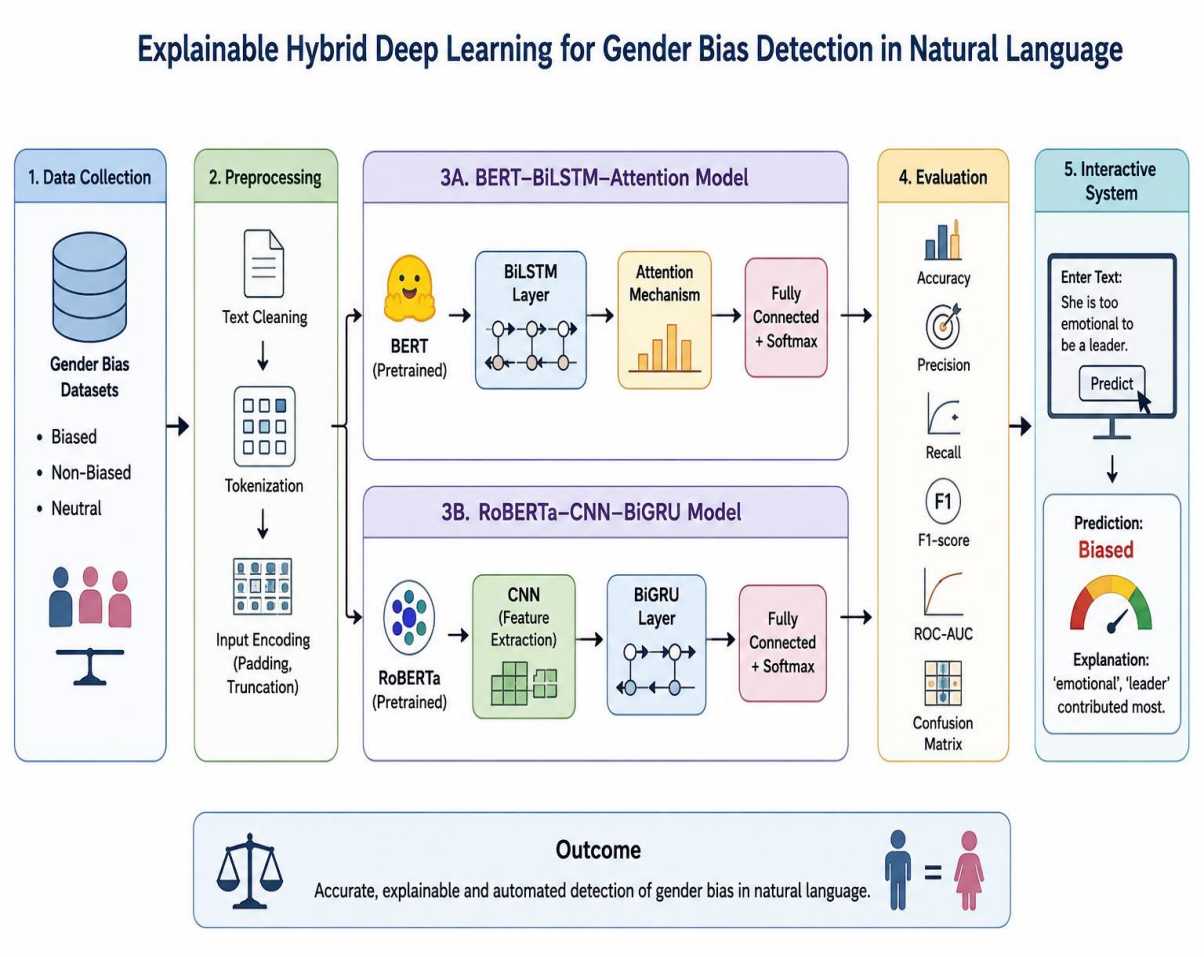

Figure 1. Research Flow Diagram

Data preprocessing is performed before model training. Text cleaning includes removing unnecessary spaces, correcting malformed entries, handling missing records, and eliminating exact duplicates. However, excessive preprocessing is avoided because transformer models rely on contextual and linguistic information. The cleaned sentences are tokenised using the corresponding pretrained BERT and RoBERTa tokenizers. Each sentence is converted into token IDs and attention masks, while padding and truncation are applied to maintain a consistent input length. The resulting numerical representations are then provided to the proposed deep learning architectures.

The first proposed architecture is a BERT–BiLSTM–Attention model. BERT is used to generate contextualised representations of the input sentence. These representations are passed to a BiLSTM layer, which captures sequential dependencies from both forward and backward directions. An attention mechanism is then applied to assign greater importance to

informative words or contextual features associated with gender bias. The resulting representation is passed through fully connected and classification layers to produce the final class prediction. The attention component also contributes to model explainability by helping identify influential linguistic information.

The second architecture is a RoBERTa–CNN–BiGRU model. RoBERTa generates contextual representations, which are processed by CNN layers to capture important local patterns and n-gram-like linguistic features. The extracted features are subsequently passed to a BiGRU layer to capture bidirectional sequential dependencies. Fully connected layers and a classification layer then generate the predicted class. Both models are trained using the training dataset and monitored using validation performance. Hyperparameters such as learning rate, batch size, sequence length, dropout, number of hidden units, and training epochs are adjusted through controlled experimentation.

The performance of both models is evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix. These measures provide a comprehensive assessment rather than relying solely on accuracy, particularly when the dataset contains unequal class distributions. Comparative analysis is performed to determine which architecture provides better predictive performance and generalisation. Finally, the best-performing model is integrated into an interactive system in which users can enter unseen sentences and receive an automated gender-bias prediction. Explainability outputs can additionally indicate the linguistic features contributing to predictions. This methodology therefore combines robust transformer-based representation learning, hybrid deep learning, explainable prediction, and practical deployment to provide a comprehensive framework for gender-bias detection.

4. Results Analysis

Table 4.1: Performance Evaluation of the Proposed BERT–BiLSTM–Attention Model

Class / Overall Measure	Precision	Recall	F1-Score	Support
Neutral	1.0000	1.0000	1.0000	36
Gender-Marked	1.0000	1.0000	1.0000	72
Accuracy	—	—	**1.0000**	108
Macro Average	1.0000	1.0000	1.0000	108
Weighted Average	1.0000	1.0000	1.0000	108
ROC-AUC	**1.0000**	—	—	108
Average Precision	**1.0000**	—	—	108

The results presented in Table 4.1 demonstrate that the proposed BERT–BiLSTM–Attention model achieved perfect classification performance on the evaluated test set of 108 samples. Both the Neutral class (36 samples) and Gender-Marked class (72 samples) achieved a precision, recall, and F1-score of 1.0000, indicating that all test instances were correctly classified without false-positive or false-negative predictions. The overall accuracy, macro average, and weighted average also reached 1.0000 (100%), demonstrating consistent performance across both classes despite their different support sizes. Furthermore, the model achieved a ROC-AUC of 1.0000 and an Average Precision of 1.0000, indicating excellent class-separation and prediction capability. These results suggest that the combination of BERT's contextual representation, BiLSTM's bidirectional sequential learning, and the Attention mechanism effectively captures linguistic patterns associated with gender-marked language. However, the perfect scores should be interpreted cautiously because the test set contains only 108 samples; therefore, evaluation on a larger, more diverse, and independently collected dataset would be necessary to establish the model's generalisation capability and rule out potential data leakage or dataset-specific effects.

Table 4.2: Class-Wise and Overall Performance Evaluation of the Proposed RoBERTa–CNN–BiGRU Model

Class	Precision	Recall	F1-Score	Support
Non-Bias	0.9921	0.9743	0.9831	778
Gender-Bias	0.9747	0.9923	0.9834	778
Accuracy	—	—	**0.9833**	**1,556**
Macro Average	**0.9834**	**0.9833**	**0.9833**	**1,556**
Weighted Average	**0.9834**	**0.9833**	**0.9833**	**1,556**

The results in Table 4.2 demonstrate that the proposed RoBERTa–CNN–BiGRU model achieved an overall accuracy of 98.33% on the final test set containing 1,556 samples. The model performed strongly for both classes, with the Non-Bias class achieving a precision of 0.9921, recall of 0.9743, and F1-score of 0.9831, while the Gender-Bias class achieved a precision of 0.9747, recall of 0.9923, and F1-score of 0.9834. The slightly higher recall for the Gender-Bias class indicates that the model was particularly effective at identifying biased instances, correctly detecting approximately 99.23% of the gender-bias samples. The macro and weighted averages are both approximately 0.9833–0.9834, showing that the model maintains balanced performance across the two equally represented classes. These results indicate that the combination of RoBERTa's contextual language representations, CNN-based local feature extraction, and BiGRU-based bidirectional sequence learning is highly effective for distinguishing biased and non-biased language. The small difference between precision and recall also suggests that the model produces relatively few false-positive and false-negative predictions. Overall, the findings confirm the effectiveness of the proposed hybrid architecture for automated gender-bias detection and provide a strong basis for comparison with the BERT–BiLSTM–Attention model.

5. CONCLUSION AND FUTURE WORK

This research presented an automated and explainable framework for detecting gender bias in natural language using hybrid deep learning architectures. The study addressed the growing concern that gender stereotypes and gender-marked expressions can be learned and reproduced by modern Natural Language Processing and Large Language Model systems. Two proposed models were investigated: the BERT–BiLSTM–Attention model and the RoBERTa–CNN–BiGRU model. The BERT–BiLSTM–Attention model achieved 100% accuracy, with precision, recall, and F1-score of 1.0000 for both the Neutral and Gender-Marked classes on the evaluated test set of 108 samples. The RoBERTa–CNN–BiGRU model also demonstrated highly competitive performance, achieving 98.33% accuracy on 1,556 test samples, with an overall F1-score of approximately 98.33%. Its balanced performance across the Non-Bias and Gender-Bias classes demonstrates its effectiveness in distinguishing biased from non-biased language.

The findings indicate that combining pretrained transformer models with sequential and feature-extraction networks can provide highly effective solutions for gender-bias classification. BERT and RoBERTa provide rich contextual representations, while BiLSTM and BiGRU capture bidirectional linguistic dependencies and CNN identifies important local textual patterns. The Attention mechanism further supports the interpretability of the BERT-based architecture by highlighting influential contextual information. The development of an interactive prediction system also extends the research from experimental model development toward practical application, allowing unseen sentences to be analysed automatically. Overall, the research demonstrates that hybrid transformer-based architectures can provide a strong foundation for automated gender-bias assessment and can contribute to the broader goal of developing more transparent, fair, and responsible AI systems.

FUTURE WORK

Future research can extend this work by evaluating the proposed models on larger, more diverse, multilingual, and independently collected datasets. Although the current models achieved very high performance, particularly the perfect score obtained by the BERT–BiLSTM–Attention model, evaluation on additional datasets is necessary to determine whether the results generalise beyond the current test data. Future datasets could include

social media posts, online news, workplace communication, educational content, and LLM-generated text. Greater attention should also be given to subtle, implicit, intersectional, and context-dependent forms of gender bias that may be difficult to identify from individual sentences.

Further work can investigate advanced Explainable AI (XAI) techniques such as SHAP, LIME, Integrated Gradients, and attention-based visualisation to provide more detailed explanations of model decisions. The models can also be extended to multilingual gender-bias detection so that linguistic and cultural differences in gender representation can be examined. Future research could additionally investigate bias mitigation by applying data balancing, counterfactual data augmentation, gender-swapping techniques, adversarial learning, and debiasing during model fine-tuning. Ensemble approaches combining the strengths of BERT–BiLSTM–Attention and RoBERTa–CNN–BiGRU could also be explored to improve robustness. Finally, the interactive system can be developed into a real-time fairness auditing platform capable of analysing large collections of text and LLM outputs, providing bias scores, explanations, confidence levels, and visual reports. Such developments would strengthen the practical applicability of the proposed framework and support future research into fair, explainable, and responsible Large Language Models.

REFERENCES

1. Blodgett, S.L., Barocas, S., Daumé III, H. and Wallach, H. (2020) ‘Language (Technology) is Power: A Critical Survey of “Bias” in NLP’, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476. doi: 10.18653/v1/2020.acl-main.485.
2. Bryskina, V., Songailaitė, M. and Mandravickaitė, J. (2026) ‘Evaluating Bias Detection in Lightweight LLMs’, *Vilnius University Open Series*. doi: 10.15388/LMITT.2026.4.
3. Caliskan, A., Bryson, J.J. and Narayanan, A. (2017) ‘Semantics derived automatically from language corpora contain human-like biases’, *Science*, 356(6334), pp. 183–186. doi: 10.1126/science.aal4230.
4. Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

5. Donald, A., Galanopoulos, A., Ojha, A.K., Curry, E., Muñoz, E., Ullah, I., McCrae, J.P., Kalra, M., Saxena, S. and Iqbal, T. (2026) 'Investigating transformer models for textual bias detection in model, data, and dataspace cards', *AI and Ethics*, 6, Article 118.
6. Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R. and Ahmed, N.K. (2024) 'Bias and Fairness in Large Language Models: A Survey', *Computational Linguistics*, 50(3), pp. 1097–1179. doi: 10.1162/coli_a_00524.
7. Liu, Y. and Li, Z. (2026) 'A systematic pipeline for diagnosing and reducing gender-stereotype bias in Japanese PLMs for sentiment analysis', *Frontiers in Artificial Intelligence*, 9, 1851069. doi: 10.3389/frai.2026.1851069.
8. Li, K., Yan, R. and Dong, Y. (2026) 'HierBias: Context-Conditioned Hierarchical Media Bias Detection with Multi-Task Type Classification', *arXiv*. doi: 10.48550/arXiv.2606.26100.
9. Mohammadi, H., Wang, S., Raeissi, M.M. and Giachanou, A. (2026) 'Reliability-Aware Sexism Detection: Combining DPO with Annotator Agreement and Token-Level Confidence Scoring', *arXiv*. doi: 10.48550/arXiv.2608.12330.
10. Nangia, N., Vania, C., Bhalerao, R. and Bowman, S.R. (2020) 'CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models', *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1953–1967. doi: 10.18653/v1/2020.emnlp-main.154.
11. Navigli, R., Conia, S. and Ross, B. (2023) 'Biases in Large Language Models: Origins, Inventory and Discussion', *Journal of Data and Information Quality*, 15(2), Article 10, pp. 1–21. doi: 10.1145/3597307.
12. Park, S., Cho, E., Jung, C.Y., Kang, W.C., Kim, T., Park, E. and Song, S. (2026) 'A benchmark dataset for evaluating gender sensitivity in Korean political discourse with large language models', *Scientific Data*, 13, Article 32. doi: 10.1038/s41597-025-06344-3.
13. Sabir, A., Kängsepp, M. and Sharma, R. (2026) 'The Confidence Trap: Gender Bias and Predictive Certainty in LLMs', *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(46), pp. 39173–39181. doi: 10.1609/aaai.v40i46.41265.
14. Salas-Jimenez, K., Ojeda-Trueba, S.-L., Bel-Enguix, G., Lee-Romero, E. and López-Ponce, F.F. (2026) 'An LLM-based methodology for the automatic detection of bias in

the DuoWikiBias corpus', *Frontiers in Artificial Intelligence*, 9, 1791624. doi: 10.3389/frai.2026.1791624.

15. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L.A., Isaac, W., Legassick, S., Irving, G. and Gabriel, I. (2022) 'Taxonomy of risks posed by language models', *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229. doi: 10.1145/3531146.3533088.
16. Liu, Y., et al. (2019) 'RoBERTa: A Robustly Optimized BERT Pretraining Approach', *arXiv*. doi: 10.48550/arXiv.1907.11692.
17. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. and Le, Q.V. (2019) 'XLNet: Generalized Autoregressive Pretraining for Language Understanding', *Advances in Neural Information Processing Systems*, 32.
18. Clark, K., Luong, M.-T., Le, Q.V. and Manning, C.D. (2020) 'ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators', *International Conference on Learning Representations (ICLR)*.
19. Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2019) 'DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter', *arXiv*. doi: 10.48550/arXiv.1910.01108.
20. Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Yu, T., Deilamsalehy, H., Zhang, R., Kim, S. and Dernoncourt, F. (2024) 'Self-Debiasing Large Language Models: Zero-Shot Recognition and Reduction of Stereotypes', *arXiv*. doi: 10.48550/arXiv.2402.01981.