Multi-Objective Optimization for Gender Bias Mitigation in Word Embeddings Applied to Hate Speech Detection

Gustavo Augusto Pires
Bachelor's Degree in Systems Engineering
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
gustavoaugustopires@ufmg.br

Luisa Marques Laboissiere
Bachelor's Degree in Systems Engineering
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
luisaml@ufmg.br

João Pedro A.F. Campos
Graduate Program in Electrical Engineering
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
jpafcampos@ufmg.br

Michel Bessani
Department of Electrical Engineering
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
mbessani@ufmg.br

***Abstract*—Word embeddings have become core components in modern natural language processing (NLP) systems, because they are important tools to represent words as a set of vectors. However, these representations often reflect and perpetuate societal biases from the training data, particularly gender bias. To address this issue, recent research has proposed multi-objective optimization strategies that aim to mitigate bias while preserving the semantic integrity of the embeddings. In this study, we investigate the practical impact of such optimized embeddings in the context of hate speech detection. This application is one of the most socially sensitive NLP tasks. We apply embeddings optimized for fairness and semantic correlation to two benchmark datasets, training neural classifiers under identical conditions. Our results show that bias-reduced and semantically-tuned embeddings maintain, and in some cases slightly improve, classification performance compared to the original representations. These findings demonstrate that it is possible to reduce gender bias in word embeddings without compromising their downstream effectiveness. This work contributes to the development of fairer and more reliable NLP systems suitable for deployment in real-world moderation and decision-making contexts.**

***Index Terms*—Gender bias, word embeddings, multi-objective optimization, hate speech detection;**

I. Introduction

Artificial intelligence (AI) systems are increasingly present in automated decision-making, influencing everything from recommendation mechanisms to critical tasks such as candidate selection, financial risk assessment, and content moderation. These systems are often embedded in platforms that reach millions of users daily, shaping access to information and opportunities. However, the growing use of language models trained on large volumes of online text has revealed the reproduction and, sometimes, amplification of historical stereotypes and social inequalities, particularly with regard to gender bias [6], [10], [11].

Gender bias, in this context, can manifest subtly, such as in the inappropriate association between professions and gender identities (e.g., linking "nurse" to women and "engineer" to men) [15], [7], or explicitly, as in the reinforcement of misogynistic language and the unequal treatment of female-coded terms. These associations are especially harmful when language technologies are applied to high-stakes or sensitive domains, such as hate speech detection [17], [18], where they may influence which content is flagged or ignored by moderation algorithms.

One of the main technical mechanisms through which this bias is internalized in systems is through linguistic representations known as word embeddings. These vector models, which capture semantic relationships between words based on their co-occurrence in large corpora, also encode the discriminatory patterns present in the source data [20], [14], [16]. Studies show that widely used embeddings, such as Word2Vec and GloVe, associate terms like "engineer" with male figures and "nurse" with female figures [7], [9]. These biases are not merely technical artifacts; when propagated into downstream applications, they risk entrenching discriminatory assumptions at scale.

Several approaches have been proposed to mitigate bias in embeddings, including pre-processing techniques, adjustments in training algorithms, and post-processing methods such as Hard Debiasing and Linear Projection [8], [7]. However, these techniques often prioritize only the removal of bias at the expense of the semantic integrity of the models, potentially impairing their usefulness in real applications. Moreover,

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), and Fundação de Amparo à Pesquisa do Estado de Minas Gerais – FAPEMIG.

they do not always consider the trade-offs between fairness and linguistic expressiveness, limiting their effectiveness in practice.

In this context, [9] proposed a solution based on multi-objective optimization, simultaneously optimizing gender fairness and the semantic quality of embeddings. This proposal represents an important advance as it does not sacrifice the linguistic effectiveness of the model in the name of equity. Instead, it explicitly models the tension between bias reduction and semantic preservation by adopting two conflicting objectives, generating a Pareto front of optimized embedding vectors.

In this work, we explore the practical application of embeddings optimized by an evolutionary multi-objective search in a real NLP task: hate speech detection. Our goal is to assess whether the improvements in fairness promoted by debiasing methods remain effective in complex and sensitive scenarios such as this, where the analyzed content often exhibits explicit or implicit bias against women and other gender minorities [12], [17], [18]. The analysis includes comparing original and optimized embeddings in supervised classifiers, aiming to understand the viability of using these models in real content moderation applications.

This article is organized as follows: Section II presents the related works, covering the main approaches to gender bias mitigation in word embeddings. Section III describes the methodology used for optimizing the word embeddings, and training and evaluating the optimized embeddings in hate speech classification tasks. Section IV details the experiments conducted and discusses the results obtained in light of the proposed objectives. Finally, Section V presents the conclusions and directions for future research.

II. Related Works

The presence of gender biases in language models has been extensively documented, reflecting historical inequalities embedded in the data used to train these systems. Such biases can directly affect applications in various areas, such as sentiment classification, machine translation, and, critically, hate speech detection. The literature indicates that these distortions not only reproduce stereotypes, but also intensify them in sensitive contexts [6], [10].

In the context of word embeddings, the vector representation of terms based on semantic co-occurrences has proven vulnerable to incorporating biased associations, especially when profession-related words are closer to masculine or feminine terms [15], [20]. Studies indicate that the problem persists in both static embeddings, such as GloVe and Word2Vec, and contextualized embeddings produced by models such as BERT and T5 [10], [8]. This characteristic is particularly problematic in automated tasks that require impartiality, such as offensive content classification, where bias can lead to unfair judgments [14].

Among mitigation strategies, post-processing methods such as removal of gender subspaces and orthogonal projections have been widely explored. The Hard Debiasing method, for example, neutralizes vectors concerning previously identified gender directions, although its effectiveness is contested in some scenarios [15], [8]. Recent advances propose formulating the bias problem as a multi-objective optimization task, as a way to balance bias reduction with the semantic preservation of vectors [13]. This approach has proven effective in maintaining the quality of embeddings while promoting greater fairness, and was successfully applied in hate speech classification tasks [12].

The relevance of investigating gender biases in hate speech contexts stems from the very nature of the phenomenon, often linked to expressions of misogyny, machismo, and sexist stereotypes. In this scenario, the use of optimized embeddings that minimize biased associations can significantly contribute to fairer and less discriminatory decisions [14], [7].

III. Methods

A. Datasets

In this study, we employed two datasets to evaluate the effect of the bias optimization in binary hate speech classification models. All datasets were pre-processed to ensure a binary labeling format, distinguishing only between the presence and absence of hate speech.

The first dataset is derived from a white supremacist forum and is known as the Hate Speech Dataset from a White Supremacy Forum (HSD-WSF). It consists of thousands of English-language sentences manually annotated as either hate speech or not. The annotation process involved a custom-built tool that allowed annotators to consult the broader context of each sentence before labeling [4]. In our work, we retained only the binary label indicating the presence or absence of hate speech, disregarding contextual metadata to ensure compatibility with the other corpora used.

The second dataset, Automated Hate Speech Detection and the Problem of Offensive Language (AHSD-POL), introduced in the study presented in [3], was constructed from tweets collected using a lexicon of hate speech-related keywords and later annotated through crowd-sourcing into three categories: hate speech, offensive language, and neither. For our binary classification task, we relabeled the data to treat only those tweets explicitly labeled as hate speech as positive samples, while combining the remaining categories (“offensive language” and “neither”) into a single negative class.

B. Bias and Semantic Evaluation Metrics

To measure the gender bias in word embeddings, we used the metric proposed by Hort et al. [9], which evaluates the bias using the Word Embedding Association Test (WEAT). This method quantifies how unequally words are associated with male and female concepts by computing the difference in proximity between each target word and two sets of gender attribute words. Rather than using similarity scores directly, we work with cosine distance, which reflects how far apart two words are in the embedding space.

The cosine distance between two word vectors $\vec{a}$ and $\vec{b}$ is defined as,

$$\mathrm{d}_{\cos}(\vec{a},\vec{b}) = 1 - \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\,\|\vec{b}\|} \tag{1}$$

.

This cosine distance ranges from 0 (more related) to 2 (less related). In most word embedding models, values typically range from 0 to 1, and higher distances indicate weaker semantic associations between words.

To measure the bias of the word embedding, the method calculates the difference between the average distances of a given word, provided by WEAT, to male-associated and female-associated terms. The higher this difference, the stronger the gender asymmetry associated with that word. This process is repeated across a set of target words, and the resulting values are added together to yield a total bias score b for the embedding model.

For example, in a pre-trained GloVe model of 50 dimensions, the word "Harry" had a cosine distance of 0.5406 to the word "family" and "Elizabeth" had a smaller distance of 0.3976 to the same word. Similarly, "Harry" had a distance of 1.0044 to the word "laundry", while "Elizabeth" had a distance of 0.8979. These numbers illustrate that, in the vector space of this model, "Elizabeth" is more closely associated with family and domestic tasks than "Harry". In this way, it is possible to say that this word embedding suffers from gender bias.

In parallel, to ensure that the semantic quality of embeddings is preserved during optimization, we used the Spearman rank correlation coefficient (ρ). This metric compares the ranking of word pair similarities in the embedding space against human-annotated similarity scores. A high correlation indicates that the model captures semantic relationships faithfully, while a drop in correlation reflects semantic degradation. It is interpreted as a cost in the optimization process.

C. Multiobjective Optimization

The multiobjective optimization problem can be stated as:

$$\arg\min_{\vec{x}\in\mathcal{X}}\{f_1(\vec{x}),\ f_2(\vec{x})\} \tag{2}$$

where $\vec{x} \in \mathcal{X}$ is a solution vector of real numbers with the same length as the vector of a given word embedding model $\vec{w}$. The objective functions, $f_1(\vec{x})$ for the bias (b) and $f_2(\vec{x})$ for semantic distortion ($1-\rho$), as presented in the previous subsection, are evaluated after using $\vec{x}$ to modify the word embedding model by multiplying $\vec{w}$, element by element, by $\vec{x}$, which generates an adjusted word embedding model $\vec{w}_{adj} = \vec{w} \circ \vec{x}$.

The Non-dominated Sorting Genetic Algorithm II (NSGA-II) [1] was used to perform the multiobjective optimization. Solutions are encoded as real-valued vectors and initialized by adding a noise vector $\vec{n}$ to a unit vector $\vec{1}$. The unit vector corresponds to not changing the word embedding model. Crossover and mutation operators are two-point crossover and noise vector addition, respectively. The noise vectors are sampled from a uniform distribution $U \sim [-0.05,\ 0.05]$.

D. Classification with Optimized Embeddings

The tweets were tokenized using the BERT tokenizer, and embeddings were obtained by concatenating individual word vectors to create fixed-length representations. A neural network classifier with the following architecture was trained for both datasets:

- Input layer (vector size: 50 * 100, max tweet length: 100 words)
- Dense layer (128 neurons, ReLU activation, dropout: 0.2)
- Dense layer (64 neurons, ReLU activation, dropout: 0.2)
- Output layer (sigmoid for HSD-WSF, softmax for AHSD-POL)

The model was trained using the Adam optimizer and binary loss.

E. Evaluation Metrics

To evaluate the performance of the classifier models trained with different word embeddings, we adopted a set of performance metrics. Relying solely on classification accuracy can be misleading, particularly in tasks with class imbalance or nuanced decision boundaries, such as hate speech detection. Therefore, we incorporated multiple complementary metrics to ensure a fair and informative comparison across models. Below, we describe the metrics used:

- Accuracy represents the proportion of correct predictions among all predictions made. While it provides a general overview of model performance, it does not distinguish between the types of errors made, nor does it reflect how well the model handles minority classes [2].
- Precision measures the proportion of true positive predictions among all instances predicted as positive. In the context of hate speech detection, high precision indicates that the model is conservative in assigning the hate speech label, minimizing false accusations of harmful content [5].
- Recall or sensitivity captures the proportion of actual hate speech instances correctly identified by the model. A high recall indicates that the model effectively detects most hateful content, which is essential in applications where missing such instances can have serious social consequences [5].
- F1 Score is the harmonic mean of precision and recall. It provides a balanced measure that penalizes models that achieve high precision at the cost of low recall, or vice versa. The F1 score is particularly useful when the dataset is imbalanced or when both types of classification errors (false positives and false negatives) carry significant weight [5].
- Area Under the Receiver Operating Characteristic Curve (AUC) measures the model's ability to discriminate between the positive and negative classes across all possible decision thresholds. It reflects the model's ranking quality across all thresholds. Although it is threshold-independent, AUC can still be affected by class imbalance, especially when one class dominates the distribution

of scores. A higher AUC indicates better overall separability between the classes [19].

Together, these metrics provide a multi-faceted understanding of model behavior, allowing us to evaluate not only overall accuracy but also the reliability, robustness, and fairness of each embedding configuration when applied to hate speech classification.

IV. Results

A. Bias Optimization

The NSGA-II algorithm was used with the following parameters: Population size of 50, crossover probability of 0.6, mutation probability of 0.2, binary tournament selection, and number of generations of 30.

The Pareto frontier estimated by the NSGA-II, illustrated in Figure 1, demonstrates an improvement over the original word embeddings in both evaluated objectives: semantic similarity and gender bias. All solutions on the frontier provide lower Spearman cost, reflecting a higher semantic correlation, and most of them present lower bias when compared to the original embedding model, represented by the black diamond in Figure 1. This result suggests that multi-objective optimization can effectively reconcile fairness and performance in embedding spaces.

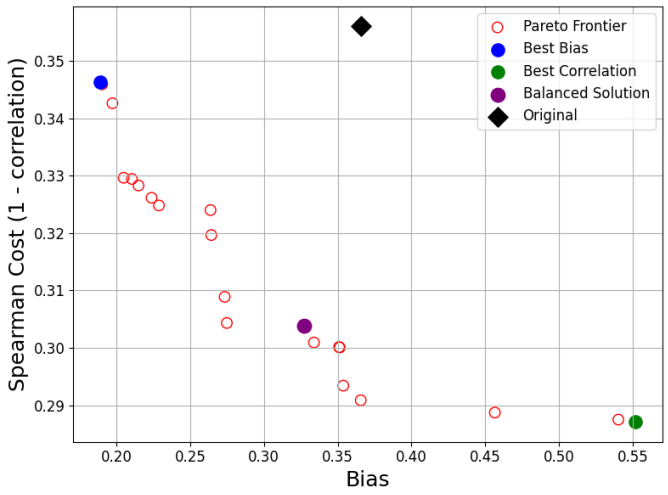

Fig. 1. Pareto frontier.

Notably, the embeddings positioned along the Pareto frontier reveal a trade-off curve: as gender bias is reduced, there is a gradual increase in semantic distortion, and vice versa. However, several solutions demonstrate substantial bias mitigation with only marginal losses in semantic alignment. This observation reinforces that bias reduction does not necessarily entail a compromise in model utility. Among the optimized candidates, three representative solutions were highlighted: the one with the lowest bias, the one with the highest semantic correlation, and a balanced solution minimizing the sum of normalized objective functions.

B. Task results

To evaluate the practical impact of bias-optimized embeddings on real-world tasks, we applied our models to two benchmark datasets: the Hate Speech Dataset from a White Supremacy Forum (HSD-WSF) and the Automated Hate Speech Detection and the Problem of Offensive Language (AHSD-POL). Each model, based on different embedding configurations, was trained using the same neural architecture described in Section III. We evaluated four embedding variants: the original embedding, the weighted solution from our multi-objective optimization, the best-performing solution in terms of semantic correlation (Spearman), and the embedding with the lowest gender bias.

Tables I and II present the comparative results for the original word embedding model and the three optimized versions across five classification metrics: accuracy, F1 score, AUC, precision, and recall.

TABLE I
CLASSIFICATION RESULTS ON THE HSD-WSF DATASET (WHITE SUPREMACY FORUM).

Embedding	Accuracy	F1 Score	AUC	Precision	Recall
Original	0.9288	0.1018	0.6431	0.1942	0.0690
Weighted	0.9260	0.1241	0.6293	0.2016	0.0897
Best Spearman	0.9288	0.1108	0.6227	0.2056	0.0759
Best Bias	0.9290	0.1373	0.6334	0.2373	0.0966

TABLE II
CLASSIFICATION RESULTS ON THE AHSD-POL DATASET (ANNOTATED TWITTER POSTS).

Embedding	Accuracy	F1 Score	AUC	Precision	Recall
Original	0.8596	0.3563	0.7976	0.4344	0.3020
Weighted	0.8599	0.3298	0.7813	0.4292	0.2678
Best Spearman	0.8398	0.3389	0.7875	0.3613	0.3191
Best Bias	0.8405	0.3192	0.7847	0.3542	0.2906

The results reveal key insights about the effect of embedding optimization. In both datasets, the models using bias-optimized or semantically-tuned embeddings perform comparably to the original embedding in terms of classification accuracy. In some cases, such as the Best Bias embedding on HSD-WSF, the alternative embeddings even outperform the original model in precision and F1 score. These numbers indicate a better balance between correctly identifying hate speech and avoiding false positives.

Importantly, the differences in performance across the four embedding variants are minor in magnitude and consistent across metrics. While there are some small trade-offs in AUC or recall, none of the optimized embeddings lead to significant degradation in the model's ability to detect hate speech.

From a broader perspective, these findings suggest that multi-objective optimization of word embeddings, aimed at reducing bias while preserving semantic quality, may not necessarily hinder classification performance. On the contrary, embeddings optimized for fairness and semantic correlation appear to maintain, and in some cases slightly improve, task effectiveness when compared to the original vectors.

This suggests that it is possible to reduce social bias in word representations without sacrificing task effectiveness. Such a result is especially relevant for deploying language models in sensitive domains, where ethical considerations are paramount. It also reinforces the value of integrating fairness objectives directly into the embedding learning or transformation process, rather than treating them as post hoc concerns.

V. Conclusion

This study investigated whether multi-objective optimization of word embeddings, targeting both gender fairness and semantic preservation, compromises performance in hate speech detection tasks. We applied embedding transformations in Word2Vec to handle with two datasets, HSD-WSF (white supremacist forum) and AHSD-POL (annotated Twitter corpus). We evaluated them using a consistent neural classifier architecture.

The results showed that embeddings optimized to reduce gender bias or maximize semantic correlation achieved classification performance largely on par with, or slightly better than, the original embeddings. In the HSD-WSF dataset, the bias-optimized embedding produced improvements in both F1 score and precision, while maintaining the highest accuracy across models. In the AHSD-POL dataset, although there was a slight decrease in F1 and recall compared to the baseline, the performance drop was minimal and did not affect overall classification effectiveness.

These findings indicate that the multi-objective debiasing process can mitigate gender bias. In practical terms, this demonstrates that fairness-aware representation learning can coexist with task efficacy, countering the belief that reducing bias necessarily entails a loss in utility.

Future research should explore scalability across more diverse languages, embedding types, and downstream tasks. In addition, user-centered evaluation and real-world deployment studies would help quantify the societal impact of these fairness-preserving models in active moderation environments.

Although this study focused on gender bias, future work should explore intersectional and race/ethnicity-related biases, using suitable target sets already available in the literature. Moreover, adding explainability elements, user-centered case studies (e.g., with human moderators), and experiments involving contextual embeddings (e.g., BERT, RoBERTa) can improve applicability. We also aim to make the preprocessing code and optimization seeds publicly available to ensure full reproducibility of results.

References

[1] Kalyanmoy Deb et al. "A fast and elitist multiobjective genetic algorithm: NSGA-II". In: *IEEE transactions on evolutionary computation* 6.2 (2002), pp. 182–197.

[2] Marina Sokolova and Guy Lapalme. "A systematic analysis of performance measures for classification tasks". In: *Information Processing & Management* 45.4 (July 2009), pp. 427–437. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2009.03.002. URL: http://dx.doi.org/10.1016/j.ipm.2009.03.002.

[3] Thomas Davidson et al. "Automated Hate Speech Detection and the Problem of Offensive Language". In: *Proceedings of the International AAAI Conference on Web and Social Media* 11.1 (May 2017), pp. 512–515. ISSN: 2162-3449. DOI: 10.1609/icwsm.v11i1.14955. URL: http://dx.doi.org/10.1609/icwsm.v11i1.14955.

[4] Ona de Gibert et al. "Hate Speech Dataset from a White Supremacy Forum". In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, 2018. DOI: 10.18653/v1/w18-5102. URL: http://dx.doi.org/10.18653/v1/W18-5102.

[5] Simon Orozco-Arias et al. "Measuring Performance Metrics of Machine Learning Algorithms for Detecting and Classifying Transposable Elements". In: *Processes* 8.6 (May 2020), p. 638. ISSN: 2227-9717. DOI: 10.3390/pr8060638. URL: http://dx.doi.org/10.3390/pr8060638.

[6] Md. Arshad Ahmed et al. "The role of biased data in computerized gender discrimination". In: *Proceedings of the Third Workshop on Gender Equality, Diversity, and Inclusion in Software Engineering*. ICSE '22. ACM, May 2022, pp. 6–11. DOI: 10.1145/3524501.3527599. URL: http://dx.doi.org/10.1145/3524501.3527599.

[7] Aylin Caliskan et al. "Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics". In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '22. ACM, July 2022, pp. 156–170. DOI: 10.1145/3514094.3534162. URL: http://dx.doi.org/10.1145/3514094.3534162.

[8] Sangeeth Ajith, M Rithani, and R S SyamDev. "Identifying and Mitigating Gender Bias in Language Models: A Fair Machine Learning Approach". In: *2023 Seventh International Conference on Image Information Processing (ICIIP)*. IEEE, Nov. 2023, pp. 888–893. DOI: 10.1109/iciip61524.2023.10537623. URL: http://dx.doi.org/10.1109/ICIIP61524.2023.10537623.

[9] Max Hort, Rebecca Moussa, and Federica Sarro. "Multi-objective search for gender-fair and semantically correct word embeddings". In: *Applied Soft Computing* 133 (Jan. 2023), p. 109916. ISSN: 1568-4946. DOI: 10.1016/j.asoc.2022.109916. URL: http://dx.doi.org/10.1016/j.asoc.2022.109916.

[10] Pradeep Kamboj, Shailender Kumar, and Vikram Goyal. "Measuring and Mitigating Gender Bias in Contextualized Word Embeddings". In: *2023 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*. IEEE, Oct. 2023, pp. 1–5. DOI: 10.

1109/icbds58040.2023.10346586. URL: http://dx.doi.org/10.1109/ICBDS58040.2023.10346586.

[11] Hadas Kotek, Rikker Dockum, and David Sun. "Gender bias and stereotypes in Large Language Models". In: *Proceedings of The ACM Collective Intelligence Conference*. CI '23. ACM, Nov. 2023, pp. 12–24. DOI: 10.1145/3582269.3615599. URL: http://dx.doi.org/10.1145/3582269.3615599.

[12] Zhi Ling. "Resolving Gendered Ambiguous Pronouns with Gender-Fair Modeling Based on BERT Word Embeddings". In: *Proceedings of the 2023 9th International Conference on Computing and Artificial Intelligence*. ICCAI 2023. ACM, Mar. 2023, pp. 523–528. DOI: 10.1145/3594315.3594367. URL: http://dx.doi.org/10.1145/3594315.3594367.

[13] Abhishek Mandal, Suzanne Little, and Susan Leavy. "Multimodal Bias: Assessing Gender Bias in Computer Vision Models with NLP Techniques". In: *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*. ICMI '23. ACM, Oct. 2023, pp. 416–424. DOI: 10.1145/3577190.3614156. URL: http://dx.doi.org/10.1145/3577190.3614156.

[14] Mustafa Bozdag, Nurullah Sevim, and Aykut Koç. "Measuring and Mitigating Gender Bias in Legal Contextualized Language Models". In: *ACM Transactions on Knowledge Discovery from Data* 18.4 (Feb. 2024), pp. 1–26. ISSN: 1556-472X. DOI: 10.1145/3628602. URL: http://dx.doi.org/10.1145/3628602.

[15] Rishabh Jain. "Assessing Gender Bias in Machine Translation". In: *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. IEEE, June 2024, pp. 1091–1096. DOI: 10.1109/icaaic60222.2024.10575537. URL: http://dx.doi.org/10.1109/ICAAIC60222.2024.10575537.

[16] Samia Kabir, Lixiang Li, and Tianyi Zhang. "STILE: Exploring and Debugging Social Biases in Pre-trained Text Representations". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI '24. ACM, May 2024, pp. 1–20. DOI: 10.1145/3613904.3642111. URL: http://dx.doi.org/10.1145/3613904.3642111.

[17] Fahim Muntasir and Jannatun Noor. "Explainable AI Discloses Gender Bias in Sexism Detection Algorithm". In: *Proceedings of the 11th International Conference on Networking, Systems, and Security*. NSysS '24. ACM, Dec. 2024, pp. 120–127. DOI: 10.1145/3704522.3704524. URL: http://dx.doi.org/10.1145/3704522.3704524.

[18] Christian Javier Ratovicius, J. Andrés Diaz-Pace, and Antonela Tommasel. "Detection of Gender Bias in Legal Texts Using Classification and LLMs". In: *2024 L Latin American Computer Conference (CLEI)*. IEEE, Aug. 2024, pp. 1–10. DOI: 10.1109/clei64178.2024.10700116. URL: http://dx.doi.org/10.1109/CLEI64178.2024.10700116.

[19] Eve Richardson et al. "The receiver operating characteristic curve accurately assesses imbalanced datasets". In: *Patterns* 5.6 (June 2024), p. 100994. ISSN: 2666-3899. DOI: 10.1016/j.patter.2024.100994. URL: http://dx.doi.org/10.1016/j.patter.2024.100994.

[20] Sarvesh Tiku. "Mitigation of User-Prompt Bias in Large Language Models: A Natural Langauge Processing and Deep Learning Based Framework". In: *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*. IEEE, Apr. 2024, pp. 1–5. DOI: 10.1109/icmi60790.2024.10585628. URL: http://dx.doi.org/10.1109/ICMI60790.2024.10585628.