Bias Detection and Mitigation in LLM-Generated Text

Aneesh Goud Mamindla

Shiva Karthik Pinjarle Manmohan

Meghna Reddi

December 16, 2024

Contents

Abstract

This report investigates the presence and implications of bias in text generated by Large Language Models (LLMs) and proposes a comprehensive framework for its detection and mitigation. With the rapid integration of LLMs into critical sectors such as hiring, healthcare, education, and legal services, the potential for perpetuating existing biases and introducing new ones is significant. This can adversely affect decision-making processes and reinforce societal inequalities. To address these challenges, our project employs a two-pronged approach: first, by using the AIF360 toolkit to quantitatively measure bias in terms of disparate impact, statistical parity difference, equal opportunity difference, average odds difference, and the Theil index; and second, by implementing adversarial debiasing techniques that include pre-processing for data balancing, in-processing for model training, and post-processing for results evaluation. Additionally, this project expands the analysis to cross-lingual contexts by translating the dataset into Hindi and reassessing bias metrics to explore the persistence of biases across languages. The findings indicate a significant reduction in bias post-mitigation, although they also reveal inherent trade-offs between fairness and predictive performance. This report contributes to the ongoing discourse on ethical AI by demonstrating effective strategies for enhancing the fairness and transparency of AI systems, thereby paving the way for more trustworthy technology implementations.

1 Introduction

The emergence of large language models in recent times has demonstrated exceptional performance across the domains of Natural Language Processing. Such models are specifically designed to automate natural language processing tasks such as generating text that mimics the style of humans, language translation with high fluency and accuracy, and categorizing emotions captured in a given text. Using a large corpus of training data and human feedback, the models process the given input and draw relevant and coherent conclusions with the given objective. The advancements in large language models have been utilized in diverse fields, unlocking their potential to be efficient and effective in enhancing decision-making and streamlining complex tasks. However, such remarkable progress comes with downsides to it. For instance, training the models on real-world data can make them susceptible to bias from disparities and inherent patterns within the text data. It is essential to address this issue in models such as ChatGPT, as the propagation of biases can lead to the development of unfair AI systems that reinforce false stereotypes and unjust treatment toward a specific demographic.

Several methods have been introduced to tackle this critical issue of bias in language models. Techniques proposed by Bolukbasi et al. (2016) and Caliskan et al. (2017) demonstrated that word embeddings such as word2vec and GIoVe carry biases from their training data. Barikeri et al. (2021) presented REDDIT-BIAS for evaluating bias. However, mitigation preprocessing still exhibits religious biases. Nangia et al. (2020) introduced the crowdsourced Stereotype Pairs

dataset that primarily focuses on stereotypes expressed in historically disadvantaged groups. Their experiments showcase that even in advanced models such as BERT, stereotype associations were still present but less biased compared to old word embeddings. Brian et al. (2018) [29] mentioned robust methods for the training process of LLMs, such as Adversarial Learning, which uses dropout and regularization terms to handle the bias. However, the challenges posed are computationally intensive.

1.1 Objective

To detect and mitigate bias in text generated by Large Language Models (LLMs), focusing on enhancing fairness and transparency.

Our paper aims to address the issue of bias in large language models. It proposes methods to manage bias detected in the training and development phases to help produce less biased and fair AI algorithms. By incorporating fairness as an essential part of the machine learning life cycle, we hope such methods allow us to build AI systems that uphold ethical values while still performing tasks effectively. Therefore, evaluating models for their accuracy and fairness, as in this work, helps understand the effects of deploying AI systems in specific contexts, encouraging the practice of inclusivity and reducing bias throughout the ML lifecycle.

1.2 Motivation

In the present world, where people are allowed to post anything they want, there is a big problem of people posting wrong information, which is a barrier to creating a safe environment when using the internet. Social media apps and the internet, in general, are full of text that can be rude and offensive, which makes it challenging to train AI and language models that use this data. This challenge requires critical detection and bias-reduction techniques that correctly find the source of unfairness in both the data used for training the model and the decisions taken by the actual model so that they do not lead to discrimination against groups of people. To avoid further propagating stereotypes and prejudice, sophisticated bias detection techniques can be applied to the data sets, and fairness restrictions can be imposed on the model structure. These attempts are critical to advancing AI and the continued proper use and development of these technologies in today's society.

2 Problem Statement

2.1 Issue

Bias in AI-generated text is not merely a technical anomaly; it represents a fundamental challenge to ethical AI development. These biases can arise from various sources, including biased training data, flawed algorithm design, or the unintended consequences of the model's interaction with real-world data. Such

biases in AI systems, especially Large Language Models (LLMs), can perpetuate stereotypes by reinforcing harmful social prejudices through their outputs. This perpetuation can subconsciously influence human decision-making in critical areas such as recruitment, legal sentencing, loan approvals, and beyond, leading to a cycle of discrimination that can be difficult to break.

2.2 Impact

The impact of bias in AI applications is profound and far-reaching. It can disproportionately affect marginalized and vulnerable groups across dimensions such as gender, race, ethnicity, and socioeconomic status. For example, if an AI system used in hiring processes is biased against certain racial groups, it could systematically deny these groups access to employment opportunities, thus perpetuating economic disparities. Similarly, biases in AI systems used in healthcare could lead to differential treatment of patients based on gender or ethnicity, potentially affecting the quality of care and health outcomes.

Moreover, the presence of bias in AI not only affects individual fairness but also erodes public trust in AI technologies. Ensuring fairness and transparency in AI systems is therefore not only a technical necessity but also a moral imperative to promote social justice and equity. Addressing these biases requires a multi-faceted approach that includes diverse and inclusive training datasets, algorithmic transparency, and ongoing monitoring to detect and mitigate biases effectively.

2.3 Societal Relevance

The societal implications of biased AI are significant, as these systems are increasingly integrated into everyday life. The decisions made by biased AI models can have long-lasting effects on the social fabric, influencing public policy, social mobility, and the fairness of societal institutions. Therefore, the challenge of mitigating bias in AI is not only about improving technology but also about shaping the kind of society we want to live in. It requires collaborative efforts among technologists, policymakers, and community stakeholders to ensure that AI serves the public good and enhances human welfare without discrimination.

3 Methodology

3.1 Bias Detection Framework

The AI Fairness 360 (AIF360) toolkit, developed by IBM, represents a robust framework utilized in this project for detecting and mitigating bias in AI-generated text. This toolkit provides an extensible, open-source library featuring a comprehensive suite of metrics and algorithms designed specifically to identify, understand, and rectify discriminatory biases that may exist within datasets and machine learning models.

3.1.1 Metrics Employed

AIF360 employs a variety of fairness metrics to quantitatively assess the presence of bias in AI models, including but not limited to:

- **Disparate Impact**: This metric compares the probability of favorable outcomes between unprivileged and privileged groups. A value of 1 indicates perfect fairness, while values less than 1 suggest bias against the unprivileged group.
- **Statistical Parity Difference**: Measures the difference in the probability of receiving favorable outcomes between the privileged and unprivileged groups. Closer to zero means less bias.
- **Equal Opportunity Difference**: Focuses on the difference in true positive rates between unprivileged and privileged groups. It helps to ensure that all groups have equal chances of receiving positive outcomes when they should.
- **Average Odds Difference**: Averages the difference in false positive rates and true positive rates between unprivileged and privileged groups, providing a more comprehensive view of bias in outcomes.

Each of these metrics offers a unique lens through which to view the fairness of a model, allowing developers to identify specific types of bias and take appropriate corrective actions.

3.1.2 Algorithmic Solutions

AIF360 not only identifies bias but also provides algorithms to mitigate it, which can be categorized into pre-processing, in-processing, and post-processing approaches:

- **Pre-processing Algorithms**: These algorithms work by modifying the training data to remove biases before model training. For example, reweighting instances in the data so that the classifier does not learn the existing biases.
- **In-processing Algorithms**: These involve modifying the learning algorithm itself to reduce bias during the training of the model. For example, adversarial debiasing, which pits a classifier against an adversary that seeks to discover bias in the classifier's predictions.
- **Post-processing Algorithms**: Applied after the model has been trained, these algorithms adjust the model's output to ensure fairness. This can include altering the decision thresholds for different groups to equalize outcomes.

3.1.3 Overall Impact

The integration of the AIF360 toolkit into this project enhances our capability to rigorously analyze and address biases. By leveraging its comprehensive set of metrics and algorithms, we can systematically identify bias at various stages of the AI lifecycle and apply targeted interventions to create more equitable AI systems. This methodology not only improves the trustworthiness of AI applications but also aligns with broader ethical standards in AI development and deployment.

3.2 Translation and Cross-Lingual Analysis

Understanding and mitigating bias in AI-generated text requires consideration of language diversity to ensure models perform equitably across different linguistic contexts. This project extends the scope of bias detection and mitigation into a cross-lingual framework by incorporating the Hindi language, which is spoken by a significant portion of the global population. The translation process is executed using Hugging Face's state-of-the-art translation pipeline, which leverages the latest advancements in neural machine translation technology.

This methodology not only allows us to examine whether biases detected in English text are also present in its Hindi translation but also helps identify any new biases that the translation process might introduce. Furthermore, the analysis in Hindi allows for the evaluation of model performance and bias mitigation techniques in non-English texts, which is critical for developing AI systems that can be deployed globally.

Upon translating the dataset, the same bias metrics used in the English version—such as Disparate Impact, Statistical Parity Difference, and others—are recalculated for the Hindi dataset. This is to assess whether the adversarial debiasing techniques, which were applied to the English text, are effective across languages, or if biases manifest differently due to linguistic and cultural variations.

This cross-lingual analysis is pivotal as it highlights the robustness of debiasing methods and provides insights into the challenges and necessities of creating fair AI systems that are linguistically and culturally inclusive. It ultimately contributes to the broader goal of this research in promoting fairness and transparency in AI, regardless of language.

3.3 Bias Mitigation

To effectively address and reduce bias in AI-generated text, this project employs a holistic approach to adversarial debiasing, which includes several stages: preprocessing, in-processing, and post-processing. Each stage plays a critical role in ensuring the fairness of the model outputs.

3.3.1 Pre-processing: Balancing Datasets Through Oversampling

The pre-processing stage focuses on preparing the dataset to minimize existing biases before training begins. Given that many biases stem from imbalanced training data where certain groups are underrepresented, balancing the dataset is crucial. Oversampling is used to increase the frequency of underrepresented classes to a level comparable to their counterparts. This approach helps to ensure that the model is exposed to a more equitable distribution of data during training, thereby reducing the initial bias in the dataset.

3.3.2 In-processing: Training with Adversarial Debiasing

In-processing involves directly incorporating bias mitigation during the model training phase. Adversarial debiasing is a technique where a predictor model and an adversary model are trained simultaneously. The predictor model is trained to perform the main task (e.g., text classification), while the adversary model attempts to predict sensitive attributes (e.g., gender or race) from the predictor's outputs. The objective is to train the predictor in such a way that its outputs cannot be used by the adversary to predict sensitive attributes accurately, thereby ensuring that these attributes do not influence the main task predictions. This method helps in significantly reducing the bias during model training.

3.3.3 Post-processing: Evaluating Fairness Metrics

After the model is trained, post-processing involves evaluating its outputs using various fairness metrics such as Disparate Impact, Statistical Parity Difference, and others. This stage is critical to assess the effectiveness of the debiasing techniques implemented in the earlier stages. It also involves making adjustments to the model's outputs, if necessary, to ensure they meet the fairness criteria established. For instance, thresholds may be adjusted for different groups to equalize the rates of positive outcomes, thereby compensating for any residual bias that was not fully eliminated during training.

3.3.4 Overall Implications

These methods of adversarial debiasing provide a robust framework for reducing bias in AI models. However, it is essential to continuously monitor and update the methodologies as new data and scenarios emerge. Bias in AI is a dynamic issue; thus, the mitigation strategies must also be adaptable and continually evolving. Employing these comprehensive debiasing strategies helps in advancing towards the development of AI systems that are not only effective but also fair and trustworthy.

4 Results

4.1 Bias Metrics Before Translation

This section discusses the results obtained from the analysis of bias metrics before the application of debiasing techniques. The primary metrics analyzed include Disparate Impact, Statistical Parity Difference, Equal Opportunity Difference, Average Odds Difference, and the Theil Index.

- **Disparate Impact**: Initially measured at 0.177, indicating a significant disparity in outcomes between the privileged and underprivileged groups, with values less than 1 suggesting bias against the underprivileged group.
- **Statistical Parity Difference**: Started at -0.641, demonstrating a notable difference in the rate of favorable outcomes received by different demographic groups, with negative values indicating bias against the underprivileged group.
- **Equal Opportunity Difference**: Observed at 0.0, suggesting no disparity in the true positive rates between different groups.
- **Average Odds Difference**: Recorded at 0.5, showing that the average rate of both false positives and true positives is unequal across groups.
- **Theil Index**: Calculated at 0.059, measuring the entropy in prediction distribution across groups, indicating relatively low inequality.

Post debiasing, improvements were observed:

- Disparate Impact improved to 1.282, approaching parity, which suggests reduced bias.
- Statistical Parity Difference increased to 0.220, moving closer to zero and indicating less bias.

4.2 Bias Metrics After Translation

After translating the dataset into Hindi, the same metrics were recalculated to determine if the debiasing techniques maintained their effectiveness in a different linguistic context. The results were as follows:

- **Disparate Impact**: Slightly changed to 1.272 post-debiasing, maintaining near parity as in the original language, indicating effective mitigation of bias across languages.
- **Statistical Parity Difference**: Improved from -0.641 to 0.213, which, similar to the English dataset, suggests a reduction in bias post-debiasing.
- **Equal Opportunity Difference**: Remained at 0.0, consistently showing no bias in true positive rates between groups.

- **Average Odds Difference**: Stayed relatively stable at 0.485, indicating a sustained balance in the rates of false and true positives.
- **Theil Index**: Consistently low at 0.059, confirming that the translation did not introduce additional disparities in prediction distribution.

These results demonstrate that the methodologies employed are capable of reducing bias in LLM-generated text and that these techniques can be effective across different languages, highlighting the robustness and adaptability of the debiasing processes.

5 Analysis and Discussion

This section evaluates the implications of the observed results, focusing on the significant reduction in bias post-debiasing, the trade-offs encountered such as reduced accuracy, and the robustness of debiasing methods across different languages.

5.1 Reduction in Bias

The data demonstrate a clear reduction in several key bias metrics after applying debiasing techniques. For instance, the Disparate Impact metric, which ideally should be close to 1, improved markedly from 0.177 to over 1.2 in both the original and translated datasets. This indicates a movement towards equal treatment of all demographic groups by the model. The improvement in Statistical Parity Difference, from a large negative value towards zero, further confirms that the likelihood of favorable outcomes is becoming more uniform across different groups, thereby enhancing fairness.

5.2 Trade-offs Between Fairness and Performance

One of the critical observations from this study is the trade-off between increased fairness and reduced model performance. After debiasing, the accuracy of the model dropped to 57

5.3 Robustness Across Languages

The robustness of debiasing methods when applied to a translated dataset—specifically from English to Hindi—indicates that these techniques can be effectively utilized across different linguistic contexts without losing their efficacy. The consistency in bias metrics before and after translation suggests that the translation process managed to preserve the essential semantic structures of the dataset necessary for fairness analysis. However, this robustness also calls for further scrutiny, particularly regarding how translation quality could impact bias assessment and mitigation.

5.4 Implications for AI System Deployment

The findings of this study are critical for the development and deployment of AI systems in multilingual and multicultural contexts. By demonstrating that debiasing techniques can be effective across languages, the study supports the argument for designing AI systems that are inherently fair, irrespective of the linguistic or demographic context in which they are deployed. This is particularly relevant in global applications, where AI systems must perform equitably across diverse user bases.

5.5 Limitations and Future Work

While the results are promising, the limitations of the current study should be acknowledged. The dataset size and focus on only gender bias are significant constraints that might affect the generalizability of the findings. Future research should aim to include larger, more diverse datasets that cover a broader range of bias types and more languages. This would help in validating the robustness of the findings and in exploring whether similar trade-offs between fairness and accuracy are observed across different contexts and bias types.

6 Conclusion

This project has successfully demonstrated the application of advanced debiasing techniques to significantly improve fairness metrics in text generated by Large Language Models (LLMs). Through a meticulous analysis using various fairness metrics, such as Disparate Impact and Statistical Parity Difference, the project has shown that it is possible to detect and mitigate biases effectively, thus contributing to the development of more ethical and equitable AI systems.

However, the study also highlighted the inherent challenges in balancing fairness with predictive performance. The reduction in model accuracy post-debiasing underscores a prevalent issue in AI fairness—enhancing fairness often compromises the model's utility, particularly in complex decision-making scenarios. This trade-off poses a critical question for AI developers and users about the acceptable balance between ethical obligations and functional efficacy.

Looking forward, the project identifies several avenues for further research and development:

- **Expanding Datasets:** Future studies should aim to incorporate larger and more diverse datasets that capture a broader spectrum of biases across different demographic and linguistic contexts. This expansion is crucial for examining the generalizability of the current findings and for understanding the dynamics of multiple intersecting biases (e.g., race, age, disability) in AI-generated text.
- **Multi-Bias Analysis:** Exploring the mitigation of multiple types of biases simultaneously will be essential for developing comprehensive debiasing solutions that do not inadvertently prioritize one form of fairness over

another. Multi-bias analysis will also help in uncovering hidden or subtle biases that may not be apparent when examining biases individually.

- **Technological Improvements:** There is a need for continuous technological advancements in debiasing methods that can preserve or even enhance the performance of AI models while improving fairness. Innovative approaches, such as more sophisticated machine learning algorithms and better quality translation tools, should be explored.

- **Real-World Applications:** Applying these methodologies to real-world scenarios, such as hiring practices, credit scoring, and legal decision-making, could provide practical insights into the operational challenges and societal impacts of AI fairness interventions.

In conclusion, while the project lays a strong foundation for mitigating bias in LLMs, the complexities of AI fairness require sustained interdisciplinary efforts. The continuation of this research is imperative for advancing our understanding of ethical AI and for ensuring that future technologies foster an inclusive society.

References

[1] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion*. Addison-Wesley, Reading, Massachusetts, 1993.

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `https://www.fairmlbook.org`

[3] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. arXiv preprint arXiv:1810.01943, 2018.

[4] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*. Advances in Neural Information Processing Systems, 2016.

[5] Xiaolei Huang, Michael C. Dorneich, and Stephen B. Gilbert. *Examining the Impact of Cross-lingual Transfer Learning in Multilingual Models for Polylingual Bias Mitigation*. Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 2022.

[6] Natalie Parde and Rodney D. Nielsen. *Bias in AI: A Survey on the Identification and Mitigation of Bias in AI Systems*. Artificial Intelligence Review, 2022.

[7] Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. *The Impact of Translation on Sentiment Bias*. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021.

[8] Anna Jobin, Marcello Ienca, and Effy Vayena. *The global landscape of AI ethics guidelines*. Nature Machine Intelligence, volume 1, pages 389–399, 2019.