Addressing Bias in Natural Language Processing: A Comprehensive Survey

Anonymous ACL submission

Abstract

[1] Natural Language Processing (NLP) is pivotal in modern technology, powering everything from voice assistants to automated translators. However, NLP technologies are not without challenges, notably the inherent biases that can perpetuate societal stereotypes and affect critical aspects of daily life, such as job recruitment and social media interactions. This survey explores key studies, methodologies, and diverse approaches to bias mitigation in NLP. It highlights the major tasks, datasets, methods, and evaluation techniques used in the field. By providing a comprehensive overview, this work aims to inform both new and established researchers about how NLP is addressing one of its most significant challenges—ensuring language technologies are equitable and beneficial for all users.

1 Introduction

Natural Language Processing (NLP) has become an indispensable tool in modern technology, bridging human communication with machine understanding across various applications from smart assistants to automated content moderation. However, as the deployment of NLP technologies grows, so does the concern for biases embedded within them. These biases can skew outcomes in ways that reinforce societal stereotypes, leading to inequities in job recruitment, social media engagement, and beyond. This introduction sets the stage for a comprehensive survey that explores the trajectory of bias recognition and mitigation in NLP. It traces the field's progression from initial identification of biases within algorithms to the development of sophisticated techniques aimed at neutralizing these biases. This survey underscores the need for ongoing vigilance and innovation to ensure that NLP technologies remain tools for positive transformation, embodying ethical practices that promote fairness and inclusivity.

2 Historical Context and Evolution

The exploration of bias in Natural Language Processing (NLP) has been profoundly influenced by pivotal studies that established the foundations for recognizing and addressing biases within language models and their datasets. Among the earliest was the work by Bolukbasi et al. (2016), titled "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." This research identified gender biases embedded in word embeddings and highlighted how these biases could be perpetuated by algorithmic models. Their proposal for mathematically adjusting embeddings to diminish gender bias not only heightened awareness but also paved the way for subsequent research in debiasing language models.

Following this, Zhao et al. (2017) broadened the bias dialogue from static to dynamic language processing with their study, "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints." This work showed how biases could be intensified during the machine learning training phases and introduced corpus-level constraints as a technique to counteract such bias magnification, representing a significant progress in proactive bias reduction in NLP systems.

3 Addressing Harms in NLP

In the field of Natural Language Processing (NLP), biases can manifest in two primary forms: allocational harms and representational harms. Both types of bias have profound implications on fairness and equity in automated systems, impacting various social groups in different ways.

3.1 Allocational Harms

Allocational harms occur when automated systems unfairly distribute resources or opportunities among different groups, such as access to housing, credit, or parole. Studies like those by Sun et al.

[1] Word count = 1986

(2019) have reviewed the literature on mitigating gender bias across various applications, underscoring the importance of fair resource allocation. This type of harm is often less directly studied in typical NLP tasks but is a critical consideration in applications like automated hiring systems, credit scoring models, or predictive policing where NLP technologies can influence decision-making processes. The work by Der Wal et al. (2022), which examines the evolution of gender bias in language models, can indirectly influence allocational decisions by affecting how gendered language is processed and interpreted in such systems.

3.2 Representational Harms

Representational harms in NLP are observed when automated systems portray one group less favorably than another. This can range from demeaning representations to the erasure of certain groups' existence. A significant amount of recent research has focused on this type of bias. For instance, Jentzsch and Turan (2022) analyzed how sentiment analysis models like BERT can perpetuate gender biases, potentially leading to skewed representations of gender in language tasks. Similarly, Salinas et al. (2023) explored biases in the internal knowledge of large language models, revealing how these systems might embed and perpetuate racial and gender stereotypes. The work by (Bertsch et al., 2022) demonstrates how gender biases can transfer from culturally rich data sources like film scripts into NLP models, affecting the representation of genders in AI-generated content. Likewise, Limisiewicz and Marecek (2022) addressed the challenge of removing gender bias in pronoun usage without losing factual gender information, crucial for avoiding the erasure of gender identities in automated text generation. Further, Marcé and Poliak (2022) discussed the gender biases present in offensive language classification models, highlighting how these biases can lead to unfair or harmful representations of certain groups in content moderation systems. (Tal et al., 2022) added to this discussion by illustrating how larger models, despite being more accurate, often amplify stereotypes, suggesting a complex trade-off between model performance and fairness.

4 Current State of the Art in Bias Mitigation in NLP

The current advancements in bias mitigation in NLP are characterized by sophisticated techniques that utilize the latest machine learning technology across both single and multimodal systems. Particularly, deep learning approaches have been at the forefront, with recent studies focusing on the structural aspects of models like BERT and GPT to understand and mitigate bias. The work by (Leteno et al., 2024) is crucial in this context, examining how internal mechanisms of these models contribute to gender bias, shedding light on the structural changes needed to mitigate such biases.

The scope of bias mitigation has also expanded to include multimodal contexts. Given NLP being applied alongside visual and auditory data, the challenge of bias mitigation has become more complex. The study by Cabello et al. (2023) is vital in investigating biases in vision-and-language models, highlighting the necessity for bias mitigation methods that span across different forms of data to ensure overall fairness.

Furthermore, the field of fairness-aware machine learning has witnessed profound developments. This method incorporates fairness metrics directly into the objectives of machine learning models, aiming for fair outcomes across different demographic groups. Specifically, training approaches that embed demographic parity and equal opportunity are crucial for applications in sensitive areas such as recruitment and law enforcement. Havens et al. (2022) provides insights of how annotation taxonomies can be crafted to enhance inclusivity of datasets used in the systems.

Moreover, there has been a growing interest in ethical AI and sociotechnical systems. This includes developing systems with built-in capabilities to dynamically adjust behaviors to minimize potential harms, ensuring that AI technologies are not only effective but also equitable and just. This shift towards ethical AI frameworks has involved diverse stakeholders in the development and evaluation of NLP systems, promoting transparency and accountability.

5 Tackling Biases within Natural Language Processing

When it comes to mitigating biases in language models, two main approaches have emerged, ranging from technical, data-centric to holistic and sociotechnical methods. This section will explore them in detail.

5.1 Technical Debiasing Techniques

Technical debiasing focuses on the technical aspects, emphasizing on key tasks such as sentiment analysis, language translation, and coreference resolution. Research in this area often utilizes benchmark datasets that have been annotated for bias, such as the Gentered Ambiguous Pronouns (GAP) dataset Zhao et al. (2018). Additionally, efforts are made to curate balanced datasets, like Jigsaw's dataset for toxic comment classification, which is used to test bias in offensive language detection Wiegand et al. (2019).

The methods employed include modifying the loss functions to penalize biased predictions, using adversarial training to reduce model sensitivity to bias features, and applying post-processing adjustments to model outputs to balance them. For instance, (Leteno et al., 2024) explored structural modifications in BERT to mitigate biases, while (Li et al., 2022) implemented in-batch balancing techniques to ensure that retrieval models do not favor certain demographics over others.

The means of evaluation often involve measuring bias before and after the application using metrics such as Gender Bias Score for word embeddings Zhao et al. (2018), or Equality of Opportunity in classification tasks (Stacey et al., 2020). The aim of this study is to quantify reductions in bias and ensure that these reductions do not come at the cost of overall model performance.

5.2 Holistic and Sociotechnical Approaches

This approach recognizes bias as a multifaceted issue that extends beyond technical solutions to encompass societal implications. It particularly focuses on tasks that necessitate direct human interaction, such as chatbots, automated hiring systems, and social media monitoring. The need to address biases in such systems is critically examined in “A Survey of Race, Racism, and Anti-Racism in NLP” by Field et al. (2021), which likely covers the broader societal implications of bias, including racial biases that are prevalent in social media interactions. Moreover, the relevance of bias mitigation in automated hiring systems is underscored by the research conducted by Parasurama and Sedoc (2022), which explores how gendered language used in resumes can lead to biased hiring decisions. Further enriching this dialogue, Sap et al. (2019) study "The Risk of Racial Bias in Hate Speech Detection" offers insights into how NLP systems can inadvertently perpetuate racial biases when monitoring online content, illustrating the complexities of detecting hate speech without reinforcing stereotypes. Lastly, the examination of gender biases across multiple languages by Zhao et al. (2024) in "Gender Bias in Large Language Models across Multiple Languages" extends the discussion by demonstrating that biases are not confined to a single linguistic or cultural framework but are a global issue that affects multilingual NLP applications

The datasets employed in this approach are intentionally diverse and inclusive, aiming to represent a wide range of dialects, socio-economic status, and cultural backgrounds.

Take the work by Pham et al. (2024) for example, who developed a diverse tweet corpus that captures the linguistic variety of English speakers from various regions, reflecting real-world language use in digital communication. Additionally, efforts by Sap et al. (2022) to understand how annotator beliefs and identities influence toxic language detection further illustrate the complexities of creating datasets that are both comprehensive and representative.

Methods are interdisciplinary, involving teams from different academic and professional backgrounds in the development and evaluation of NLP systems. These teams integrate sociological insights into the training process, emphasizing the importance of ethically collecting and using training data. This holistic approach aims to develop NLP systems that are not only technically proficient but also socially aware and responsible. Similarly, evaluation methods are comprehensive, including user studies to assess the societal impact of NLP applications and audits conducted by third parties to ensure ethical standards are met. For example, the study by Bordia1 and Bowman (2019), which identifies and reduces gender bias in word-level language models, provides a framework for evaluating how well biases are addressed in language processing tasks. Another example is the study by (Blodgett et al., 2020), which provides a critical survey of ’Bias’ in NLP, helps frame the current discussions and methodologies in bias mitigation, highlighting the necessity for ongoing evaluation and adaptation in practices.

6 Conclusion

As we have explored throughout this survey, the journey of mitigating bias in Natural Language Processing (NLP) is both complex and critical. From early detections of biases to the development of advanced debiasing techniques, the field has made significant strides towards understanding and addressing the ethical challenges presented by automated language systems. This evolution reflects a broader shift in the discipline, moving from purely technical solutions to a more integrated approach that considers the social implications of NLP technologies. Despite the progress, the task of completely eliminating bias is ongoing and requires continued innovation and commitment. Researchers and practitioners must persist in their efforts to refine methods, expand inclusive datasets, and implement fair practices, ensuring that NLP tools not only advance in capability but also in integrity. This survey highlights the importance of multi-disciplinary collaboration and continuous evaluation to build NLP systems that are just, equitable, and truly beneficial for all segments of society.

References

Amanda Bertsch, Ashley Oh, Sanika Natu, Swetha Gangu, Alan Black, and Emma Strubell. 2022. Evaluating gender bias transfer from film data.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. *ACL Anthology*.

Tolga Bolukbasi, Kai-Weo Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Shikha Bordia1 and R. Samuel Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the NAACL*.

Laura Cabello, Emanuele Bugliarello, Stephanie Brandl, and Desmond Elliott. 2023. Evaluating bias and fairness in gender-neutral pretrained vision-and-language models.

Oskar Van Der Wal, Jaap Jumelet, Katrin Schulz, and Willem Zuidema. 2022. The birth of bias: A case study on the evolution of gender bias in an english language model.

Anjalie Field, Zeerak Waseem, Su Lin Blodgett, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in nlp.

Lucy Havens, Melissa Terras, Benjamin Bach, and Beatrice Alex. 2022. Uncertainty and inclusivity in gender bias annotation: An annotation taxonomy and annotated datasets of british english text.

F. Sophie Jentzsch and Cigdem Turan. 2022. Gender bias in bert - measuring and analysing biases through sentiment rating in a realistic downstream classification task.

Thibaud Leteno, Antoine Gourru, Charlotte Laclau, and Christophe Gravier. 2024. An investigation of structures responsible for gender bias in bert and distilbert.

Yuantong Li, Xiaokai Wei, Zijian Wang, Shen Wang, Parminder Bhatia, Xiaofei Ma, and Andrew Arnold. 2022. Debiasing neural retrieval via in-batch balancing regularization.

Tomasz Limisiewicz and David Marecek. 2022. Don't forget about pronouns: Removing gender bias in language models without losing factual gender information.

Sanjana Marcé and Adam Poliak. 2022. On gender biases in offensive language classification models.

Prasanna Parasurama and João Sedoc. 2022. Gendered language in resumes and its implications for algorithmic bias in hiring.

Nhi Pham, Lachlan Pham, and Adam Meyer. 2024. Towards better inclusivity: A diverse tweet corpus of english varieties.

Abel Salinas, Louis Penafiel, Robert McCormack, and Fred Morstatter. 2023. "i'm not racist but...": Discovering bias in the internal knowledge of large language models.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and A. Noah Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the ACL*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and A. Noah Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktaschel. 2020. Avoiding the hypothesis-only bias in natural language inference via ensemble adversarial training. In *Proceedings of EMNLP*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review.

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of EMNLP*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of NAACL*.

Jinman Zhao, Yitian Ding, Chen Jia, Yining Wang, and Zifan Qian. 2024. Gender bias in large language models across multiple languages.