The Need For Inclusive NLP: Addressing Sociodemographic Bias and Enhancing Sociotechnical Systems through Interdisciplinary Frameworks

Pranav Narayanan Venkit

Pennsylvania State University, University Park, Pennsylvania, USA
pranav.venkit@psu.edu

Abstract

Natural Language Processing systems are increasingly integrated into diverse sociotechnical contexts, playing a pivotal role in essential societal functions (Venkit et al. 2023a). These systems are employed across various domains, such as education, healthcare, and policy-making, offering technical solutions to social issues (Gautam, Venkit, and Ghosh 2024). However, a critical concern is the opaque nature of these models, often presented and utilized as 'black boxes' (O'neil 2017). Proprietary restrictions or knowledge gaps obscure the underlying interactions, making them challenging to interpret (Blodgett et al. 2020). The widespread presence of these systems necessitates a thorough examination of their inherent biases and societal implications (Blodgett et al. 2020).

Despite a growing body of research on bias in NLP, significant challenges remain. Predominantly, studies focus on race and gender biases, adopt a model-centric approach to analysis, and employ technocentric methods for bias mitigation, often neglecting the broader societal ramifications (Gupta et al. 2023). Addressing these challenges requires a more comprehensive understanding of how NLP systems, as integral components of broader sociotechnical systems, impact society.

To further explore these issues, it's important to define sociotechnical systems, which encompass the intricate interplay between social and technical components within goal-oriented activities (Venkit et al. 2023a; Cooper and Foster 1971). Although publicly available sociotechnical systems centered around NLP are widely deployed, limited research examines their societal impact (Gupta et al. 2023). This gap underscores the need for a deeper understanding of the societal implications of NLP systems. This research aims to address the complexities posed by sociodemographic biases in NLP systems and the resultant harms within these frameworks (Dev et al. 2021). Sociodemographic biases refer to systematic distortions or prejudices in data and algorithms that disproportionately affect certain demographic groups. The goal is to develop more inclusive methodologies for identifying, quantifying, and mitigating these biases, bridging the interdisciplinary divide between technical and social sciences, and proposing socially aware frameworks that address the impact of biases in these systems.

To effectively address this issue, this research is structured into three interrelated facets, each focusing on specific aspects of biases in NLP:

Facet 1: Characterization of Bias: This facet draws on a series of peer-reviewed publications that provide vital insights into recognizing biases within underrepresented minority groups, a dimension often overlooked in previous studies on NLP bias. These works advance the field by presenting methodologies for detecting and mitigating biases in NLP solutions applied in social contexts. Notably, these methods are designed to be accessible to a broad audience, eliminating the need for specialized technical expertise.

In my first study (Venkit and Wilson 2021), the focus is on public sentiment analysis and toxicity detection models, widely used in sociotechnical contexts. This research specifically examines explicit bias directed towards individuals with disabilities (PWDs), as well as race and gender biases. The second study in this facet (Venkit, Srinath, and Wilson 2022) delves into the domain of implicit bias within NLP models, addressing the subtle and often unintentional propagation of harmful stereotypes. As the field of NLP evolves from embedding-based models to large language models, concerns have emerged regarding these models' tendency to reflect a 'hegemonic' perspective, perpetuating majority-held views rather than objective truths. This is explored in my study on nationality representation and its impact on societal stereotypes (Venkit et al. 2023b). By examining LLM-generated stories for various nationalities, the research establishes a correlation between sentiment and the population of internet users in a country. These studies emphasize the importance of identifying and understanding both implicit and explicit biases in NLP models, providing critical context within the often opaque realm of public-facing models and addressing the black-box nature of these systems.

Facet 2: Comprehending the Interdisciplinary Divide: This facet takes a step back to examine the sociotechnical nature of NLP systems by exploring the potential gap between NLP research and its resonance within society. The growing disconnect between technical and social domains can lead to biases and harms directed at specific societal groups through NLP systems deployed as sociotechnical solutions. The significance of this facet lies in its capacity to illuminate the adverse consequences stemming from this divide, underscoring the need for an interdisciplinary approach rather than relying solely on a technocratic perspective. This facet involves an in-depth analysis of the interdisciplinary nature of concepts such as sentiment (Venkit et al.

2023a) and hallucinations (Venkit et al. 2024) and their application within NLP models.

Sentiment analysis, a prevalent machine learning application, influences both social and technical actors within a network. However, the definition and connotation of sentiment vary significantly across different fields, potentially leading to misunderstandings about the efficacy of such systems (Venkit et al. 2023a). My completed work in this facet (Venkit et al. 2023a, 2024) examines how disciplines such as psychology, sociology, and technology conceptualize sentiment, revealing disparities and differing applications of such shared concept within technological research. By examining literature on these shared vocabularies across various domains, the research uncovers distinct conceptualizations and highlights a lack of explicit definitions and frameworks for characterizing these technical concepts, leading to potential challenges and biases. The facet proposes ethics sheets, including critical questions, to guide practitioners in the equitable use of sentiment analysis, demonstrating the need for inclusive research endeavors to refine AI models as sociotechnical solutions.

Facet 3: Integrating Frameworks: This facet explores the complex interaction between human and AI entities, aiming for a comprehensive understanding of bias, harm, and its societal ramifications. It employs a dual-pronged approach: first, to investigate how biases in NLP systems can influence society, and second, to develop a design framework that promotes the creation of more secure and equitable NLP solutions. Grounded in actor-network theory (ANT), this research aims to establish effective development networks that encourage collaboration among developers, practitioners, and other stakeholders. The initial work within this facet (Narayanan Venkit et al. 2023) focuses on the societal impact of biased text generated by language models. This research aims to understand how biases within these models can manifest as societal harm.

A human study (Narayanan Venkit et al. 2023) is conducted to assess whether readers, unaware of the source of a news article, can detect implicit or explicit biases in the text. By employing both technical and human evaluations, the study provides a comprehensive measurement of bias, offering crucial insights into its effects and resulting harms. The interview process reveals overarching themes about the influence of bias on society, including negative knowledge transfer and the perpetuation of stereotypes. Following this, in my upcoming work, the research shifts focus to establishing effective networks, leveraging ANT principles to disseminate improved design frameworks for creating inclusive and socially conscious NLP solutions. These networks aim to serve as foundational frameworks for developing holistic solutions that avoid discriminatory pitfalls in NLP, affecting specific populations. A tangible outcome is the development of a platform and design frameworks that actively consider the perspectives and design choices of developers, diverse practitioners, and stakeholders within a sociotechnical system. By fostering the development of socially aware and less harmful technology, this research contributes significantly to the fields of NLP and AI, promoting a more equitable and inclusive technological landscape. Future research efforts will focus on creating community-centric studies to understand the various forms of bias and harm that can arise from sociotechnical systems, aiming to define and measure these phenomena more accurately.

Through these interrelated facets, this research contributes to the following:

A. Providing methodologies for detecting both explicit and implicit biases in NLP models, with a focus on underrepresented groups, such as individuals with disabilities and those affected by nationality-based biases.

B. Exploring the interdisciplinary nature of bias by investigating how concepts like sentiment and harm are understood across disciplines, promoting collaboration to address biases comprehensively.

C. Developing inclusive frameworks using ANT principles, which aim to create more equitable NLP systems by addressing the societal impacts of biased language models.

D. Ensuring responsible and ethical practices in NLP for the benefit of both technical and non-technical stakeholders.

References

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476.

Cooper, R.; and Foster, M. 1971. Sociotechnical systems. *American Psychologist*, 26(5): 467.

Dev, S.; Sheng, E.; Zhao, J.; Amstutz, A.; Sun, J.; Hou, Y.; Sanseverino, M.; Kim, J.; Nishi, A.; Peng, N.; et al. 2021. On Measures of Biases and Harms in NLP. *arXiv preprint arXiv:2108.03362*.

Gautam, S.; Venkit, P. N.; and Ghosh, S. 2024. From melting pots to misrepresentations: Exploring harms in generative ai. *arXiv preprint arXiv:2403.10776*.

Gupta, V.; Venkit, P. N.; Wilson, S.; and Passonneau, R. J. 2023. Sociodemographic Bias in Language Models: A Survey and Forward Path. *arXiv e-prints*, arXiv–2306.

Narayanan Venkit, P.; Gautam, S.; Panchanadikar, R.; Huang, T.-H.; and Wilson, S. 2023. Unmasking nationality bias: A study of human perception of nationalities in ai-generated articles. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 554–565.

O'neil, C. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Venkit, P.; Srinath, M.; Gautam, S.; Venkatraman, S.; Gupta, V.; Passonneau, R. J.; and Wilson, S. 2023a. The Sentiment Problem: A Critical Survey towards Deconstructing Sentiment Analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13743–13763.

Venkit, P. N.; Chakravorti, T.; Gupta, V.; Biggs, H.; Srinath, M.; Goswami, K.; Rajtmajer, S.; and Wilson, S. 2024. An Audit on the Perspectives and Challenges of Hallucinations in NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 6528–6548.

Venkit, P. N.; Gautam, S.; Panchanadikar, R.; Huang, T.-H.; and Wilson, S. 2023b. Nationality Bias in Text Generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 116–122.

Venkit, P. N.; Srinath, M.; and Wilson, S. 2022. A Study of Implicit Bias in Pretrained Language Models against People with Disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*, 1324–1332.

Venkit, P. N.; and Wilson, S. 2021. Identification of bias against people with disabilities in sentiment analysis and toxicity detection models. *arXiv preprint arXiv:2111.13259*.