OVERCOMING RACIAL STEREOTYPES AND PREJUDICE IN LANGUAGE

Anjalie Field, anjalief@cs.cmu.edu *Carnegie Mellon University*

Stereotypes and prejudice frequently manifest in text and can have long-term negative effects on all aspects of life, including mental health, physical health, and career trajectories [1–5]. Despite their prevalence, they are difficult to detect and mitigate, because they are often implicit and can manifest subtly, e.g., as microaggressions or condescension [6]. Natural language processing (NLP) has the power to analyze large-scale text corpora, which enables the detection of patterns that may be difficult for humans to identify in isolated incidents, but become evident from their re-occurrences in many incidents. Leveraging this technology requires developing new models capable of capturing subtle connotations indicative of prejudice in diverse types of language data. As a Data Science Fellow at Stanford, I would aim to build on my prior work on social-oriented natural language processing to (1) develop NLP models for uncovering racial stereotypes in text and (2) conduct analyses of people portrayals in high-stakes application domains and provide recommendations on how to reduce stereotyping, in collaboration with Dan Jurafsky and Jennifer Eberhardt. This work has the potential for immediate direct applications for reducing racial stereotyping in television content, police interactions, and child welfare cases due to our ongoing collaborations with practitioners in these fields. Through the Data Science Fellows program, I hope to build on my experience in fostering cross-disciplinary communities in order to contribute long-term to the establishment of data science as a new field and community.

Phase 1: Computationally modeling people portrayals Directly examining stereotypes and prejudice in text requires developing computational models capable of capturing subtle connotations aligned with social theories about affect. My prior work in this area draws from behavioral science theories that identify power, agency, and sentiment as the most important axes of affective meaning [7] and develops models to score people portrayals along these dimensions. Methodology includes integrating off-the-shelf word-level annotations with pre-trained contextualized embeddings, which leverages both prior work on developing annotated word lexicons and the abilities of modern NLP to capture context [8]. Alternative methods include directly projecting entity embeddings into constructed affective subspaces [9], and training multilingual models to infer verb connotations [10]. Figure 1 provides a finer-grained example of the type of content that these methods capture: verb choice in different languages can have subtly different connotations, and our analysis of Wikipedia articles reveals that Russian articles tend to use verbs with more negative connotations when describing LGBT people than English or Spanish articles [10].

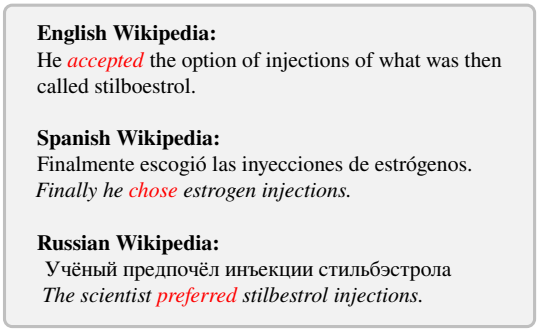

Figure 1: Alan Turing's Wikipedia page in different languages. *chose* in Spanish and *preferred* in Russian imply higher agency than *accepted* in English.

In future work, I intend to develop methods to jointly project people portrayals into different affect dimensions and examine these projections along intersected dimensions in order to target specific stereotypes. For example, the "angry Black woman" stereotype can result in negative physical, social and economic impacts, such as facilitating workplace discrimination [11, 12]. By

developing NLP methods to detect implications like emotions, we can map how people are described to vector representations in affective subspaces, which would allow us to examine targeted stereotypes. For example, we might find that "anger" has negative connotations when associated with Black women but not when associated with other demographic groups. We can develop these methods by combining existing data annotated for affect and emotions [13] with contextualized representations from large pre-trained language models [8]. I will explore models both for independently inferring scores for different affect dimensions and for jointly predicting continuous-valued vector elements. In this phase, I aim to develop generalizable methods that can easily be tailored to new domains and incorporated into analytics and interventions tools usable on web-scale data.

Phase 2: Developing recommendations based on analyses of real-word data I intend to use models for measuring and identifying stereotypes to analyze real-world data sets where practitioners are actively working to reduce stereotyping. We have identified three domains where this technology could lead to directly actionable insights. First, Profs. Jurafsky and Eberhardt are currently collaborating with BET (Black Entertainment Television) and their parent company (CBS-Viacom) to aid in creating television content that combats racial bias and promotes positive racial attitudes. My methodology for uncovering stereotypes would work towards one of the project's long-term goals: a tool that television networks and content creators can use to assess and adjust their content. Second, Profs. Jurafsky and Eberhardt additionally have an ongoing project examining the language from police body camera footage, where these methods could provide new insights into prejudice in communications [14]. Third, I am currently collaborating with the Department of Human Services in Allegheny County (DHS) on analyzing child welfare notes written by caseworkers. While our preliminary work has focused on data analysis and algorithmic fairness, one of DHS's additional priorities is understanding possible prejudices in these notes, particularly identifying signs of racial bias in order to improve the way caseworkers interact with families. This work is particularly timely, as the county is current considering deploying NLP models for information extraction over this data, which are liable absorbing and amplifying bias [15].

Analyzing stereotypes and prejudice in application domains requires integrating models developed in phase 1 with statistical analyses and methods for controlling for confounding variables. For example, child welfare agencies have different policies for children of different ages, and character portrayals might differ in T.V. shows of different genres. In each domain, I will work with domain-experts in order to understand what factors could be confounding variables that need to be controlled for or incorporated as additional analysis dimensions. I will then draw from my prior experience on using adversarial training and matching methodology to reduce the influence of confounding variables on text analyses in order to target dimensions of interest [16].

Future work Our work has the potential to develop new NLP technology capable of uncovering subtle connotations in language and have direct impact by building on our ongoing collaborations with government and industry practitioners. Longer-term, I am excited about facilitating interdisciplinary research, and I hope to develop open-source curricula, workshops, and tutorials in order to bring together researchers from different disciplines and enable sharing data and methodology. As a PhD student, I have worked towards this goal by co-organizing the ACL 2021 Workshop on NLP for Positive Impact, starting a computational and social science reading group on race and NLP in my department, and orchestrating numerous initiatives aimed at lowering access barriers and fostering diverse inclusive communities, e.g. co-coordinating graduate application support at Carnegie Mellon. As a Data Science Fellow, I would aim to form lasting collaborations that enable research in this area and contribute to the development of data science as a new field of research.

References

[1] Christine Logel, Emma C. Iserman, Paul G. Davies, Diane M. Quinn, and Steven J. Spencer. The perils of double consciousness: The role of thought suppression in stereotype threat. *Journal of Experimental Social Psychology*, 45(2):299 – 312, 2009.

[2] Claudia Goldin et al. Understanding the gender gap: An economic history of american women. *OUP Catalogue*, 1992.

[3] Nancy Krieger. Racial and gender discrimination: risk factors for high blood pressure? *Social science & medicine*, 30(12):1273–1281, 1990.

[4] David L Hamilton and Tina K Trolier. Stereotypes and stereotyping: An overview of the cognitive approach in prejudice, discrimination, and racism. 1986.

[5] Daniel Bar-Tal, Carl F Graumann, Arie W Kruglanski, and Wolfgang Stroebe. *Stereotyping and prejudice: Changing conceptions*. Springer Science & Business Media, 2013.

[6] Derald Wing Sue. *Microaggressions in everyday life: Race, gender, and sexual orientation*. John Wiley & Sons, 2010.

[7] Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. *The measurement of meaning*. Number 47. University of Illinois press, 1957.

[8] **Anjalie Field**, Gayatri Bhat, and Yulia Tsvetkov. Contextual affective analysis: A case study of people portrayals in online #MeToo stories. In *Proc. of ICWSM*, 2019.

[9] **Anjalie Field** and Yulia Tsvetkov. Entity-centric contextual affective analysis. In *Proc. of ACL*, pages 2550–2560, 2019.

[10] Chan Young Park*, Xinru Yan*, **Anjalie Field***, and Yulia Tsvetkov. Multilingual contextual affective analysis of LGBT people portrayals in Wikipedia. *Proc. of ICWSM*, pages 479–490, 2020.

[11] P.H. Collins. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Perspectives on Gender. Taylor & Francis, 1990.

[12] J. Celeste Walley-Jean. Debunking the myth of the "angry black woman": An exploration of anger in young african american women. *Black Women, Gender + Families*, 3(2):68–86, 2009.

[13] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proc. of ACL*, pages 4040–4054, Online, July 2020. Association for Computational Linguistics.

[14] Rob Voigt, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25):6521–6526, 2017.

[15] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proc. of EMNLP*, pages 2979–2989, 2017.

[16] **Anjalie Field**, Chan Young Park, and Yulia Tsvetkov. Controlled analyses of social biases in Wikipedia bios. *arXiv preprint arXiv:2101.00078*, 2020.