Rapid advancements in natural language processing (NLP) over standardized tasks threaten to reduce NLP to benchmarking metrics and overlook the ways that language fundamentally involves people. Models trained on data without considering whom it describes, whom it was written by, and whom model outputs might affect are liable to amplify stereotypes, spread misinformation, and perpetuate discrimination. I address these challenges by centering the role of people in modern NLP and developing computational approaches that identify harms to (1) people reading text [1, 2], (2) people described in text [3–7], and (3) people using and affected by NLP systems [8, 9]. These directions raise new technical challenges: much research in machine learning and AI condenses complex tasks into standardized metrics and aims to replicate human performance over reusable data sets. In contrast, my work uncovers systemic trends in large text corpora and models, tasks that are difficult for humans, involve processing complex real-world data, and cannot be achieved through supervised classification. Further, this work has numerous applications, as it addresses prominent social issues, including misinformation, racism and sexism, and child welfare. Addressing these challenges necessitates highly interdisciplinary research that includes fostering collaborations and drawing theories from related fields like causal inference, psychology and political science. Overall, I aim to promote equity, inclusion, and information integrity by developing social-oriented NLP models and providing insight into when NLP does more harm than good.

NLP models to combat text harmful to readers: bias, offensive language, and propaganda

Text can cause harms to people who read it in many ways. Hate speech and offensive language can cause distress and marginalization, and misinformation can result in physical and societal harm. Much work in NLP has aimed to address these concerns through supervised classification tasks like hate speech detection or fact checking. However, approaches relying on hand-annotated data fail to detect content that is difficult for annotators to recognize, e.g., microaggressions and propaganda. They additionally overfit to shallow lexical signals and are unable to reflect deeper pragmatics nor generalize to real-world settings. My work aims to fill this need by developing weakly supervised approaches for identifying harmful content that may be difficult for humans to detect in isolated incidents, but becomes evident from repeated patterns in large data sets.

Unsupervised models for detecting bias in conversational text Current models for offensive language detection often fail to identify subtle forms of racism or sexism, like "*you're pretty smart for a Black woman*". This content is not overtly offensive and is difficult for annotators to recognize, but it can cause harm and propagate stereotypes that are amplified in NLP models [10]. My work uncovers veiled gender bias through an unsupervised adversarially-trained model aimed to reveal implicit intents and effects: I identify systemic differences in comments addressed towards men and women by training a model to predict the gender of each comment's addressee and examining predictive features [2]. Deep learning has the capacity to process large-scale data sets and is highly adept at pattern-recognition, but current models fail to integrate causal relations, which is essential for modeling deeper pragmatic meanings. My method uses text-based propensity matching inspired by causality literature [11] and also incorporates adversarial training for demoting latent confounds [12] in order to prevent the model from focusing on spurious lexical features like names and pronouns and guide the model to learn deeper pragmatic features predictive of bias. The difficulty in controlling for overtly predictive confounds is one of the reasons that most prior work on gender bias in NLP has focused on corpus-level analyses [13, 14] even though phrase-level detection is more practically applicable.

There are numerous directions for future work in this domain, including further integration of causal inference into deep learning models, improved methods for adversarially demoting confounds, and identification of pre-training and proxy tasks for distant supervision. Furthermore, as subtle manifestations of bias are often unintentional, future research directions could focus on generating less harmful rephrasings, drawing from work on controllable text generation [15, 16].

Characterizing manipulation strategies in multilingual text While harms in text can be unintentional results of implicit bias, authors can also intentionally write biased content to manipulate readers' opinions. Known manipulation campaigns include Russian-government-affiliated accounts disingenuously tweeting about U.S. elections and rampant online misinformation about COVID-19 [17]. Common NLP approaches, e.g. fact-checking or fake news classification, fail to detect more subtle strategies, e.g. propagating content that is truthful but also polarizing or distracting.

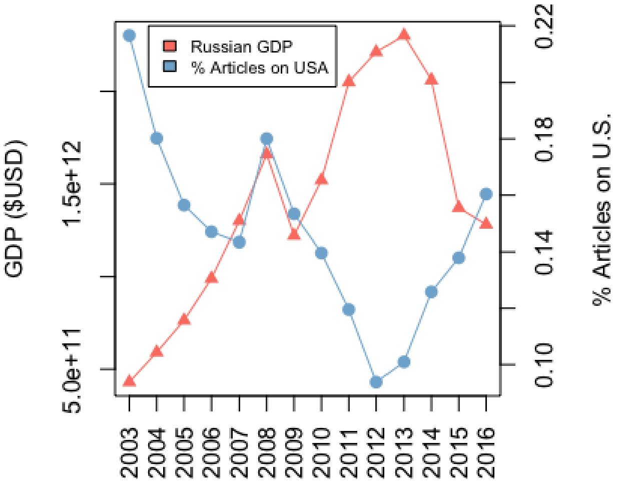

Figure 1: In [1] we uncover propaganda in Russian news by showing how Russian news coverage of the U.S. is strongly negatively correlated with the state of the Russian economy.

My work develops methods for integrating time-series analyses and network features into NLP models to uncover subtle manipulation strategies. In one domain, I analyzed a corpus of Russian newspaper articles by drawing two propaganda theories, framing and agenda-setting, from political science [18, 19]. In order to measure these abstract concepts, I incorporated Granger causality to compare news coverage with economic indicators (agenda-setting), and I developed a distant supervision approach that involved cross-lingually projecting English framing annotations into Russian. This analysis revealed how articles discuss negative events in the U.S. as a way of distracting public opinion from economic downturns in Russia (Figure 1). In a second domain, I collaborated with network scientists to develop a graph-based label propagation method for analyzing a corpus of mixed-language tweets posted about a terrorist attack in India in 2019 [20]. Both projects involved significant technical challenges over non-English data and broad open-ended research questions that were difficult to address through data annotations. Instead, I developed distantly-supervised methods for categorizing text content and grounding models in external events.

While this work sheds new light on manipulation strategies, future research is needed to understand their effects on readers: for example, when newspapers discuss moral failings of the U.S. during economic downturns, do readers also focus on the U.S. more than the economy? We can estimate some effects by examining corresponding data sources, including social media posts and surveys. Furthermore, the growth of digitized data and collection efforts has resulted in much available data from manipulation campaigns, such as U.S. campaign emails from the 2020 election [21], which NLP tools could aid in analyzing. These directions are not possible without the development of new NLP technology that can, for example, model topic and framing differences between social media and mainstream media, integrate network features to capture information spread in text, and jointly process text with GIFs, memes, and other visual data. Finally, with the growth of NLP models that generate highly fluent text, like GPT-3, more research is needed to understand what harms models absorb from disingenious training data and how to mitigate them.

NLP models for uncovering how people are described in text: stereotypes and prejudice

Text describing people is liable to perpetuating bias, stereotypes, and prejudice [22, 23], which can impact all aspects of life, including mental health, physical health, and career trajectories [24–27]. Furthermore, NLP models are prone to absorbing and amplifying biases in their training data, for example incorrectly associating female pronouns like "she" with nurses more than doctors [28, 29]. While substantial work has examined ways that models exhibit stereotypes, less research has looked further up the pipeline and examined where model biases originate – how are people described text? This broad research question requires developing computational models capable of capturing subtle connotations aligned with social theories about affect and stereotypes. In addition to having implications for developing more fair NLP models, this line of work has many real-world applications. In domains like newspaper articles, encyclopedias, or reviews of job candidates, authors often strive for objectivity, but implicit biases can result in unintentional stereotypes or prejudice.

English Wikipedia:
He *accepted* the option of injections of what was then called stilboestrol.

Spanish Wikipedia:
Finalmente escogió las inyecciones de estrógenos.
Finally he chose estrogen injections.

Russian Wikipedia:
Учёный предпочёл инъекции стильбэстрола
The scientist preferred stilbestrol injections.

Figure 2: Example from Alan Turing's Wikipedia page in different languages. The English edition uses the verb *accepted*, which suggests that Turing had little control over the situation (low agency). In contrast, *chose* in Spanish and *preferred* in Russian imply he actively made the decision (high agency).

Computationally modeling people portrayals along dimensions of power, agency and sentiment My approach to modeling how people are described in text draws from behavioral science theories that have identified power, agency, and sentiment as the most important axes of affective meaning [30] 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 [4]. Alternative methods include constructing affective subspaces and directly projecting entity embeddings into target dimensions [3], as well as training multilingual models to infer verb connotations [5]. I have additionally been working to develop high-dimensional matching approaches inspired by causal inference methods to target dimensions of interest [6]. These models have ultimately revealed signs of bias and prejudice. For example, even though the #MeToo movement has been viewed as empowering, women are often portrayed as less powerful than men in media coverage of events [4]. Figure 2 provides a finer-grained example: 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 [5].

The real-world implications of our research have lead to media coverage, including a collaboration with Washington Post analysts on examining anti-Black racism in China, and there is additionally interest in implementing our methods for analyzing Wikipedia articles at the Wikimedia Foundation.[1] Furthermore, while my prior work has focused on detecting stereotypes, NLP

[1] https://www.post-gazette.com/news/health/2019/09/01/Computational-gender-bias-MeToo-Carnegie-CMU-Ansari-media/stories/201908230135;https://www.washingtonpost.com/politics/2020/06/18/video-evidence-anti-black-discrimination-china-over-coronavirus-fears/;https://phabricator.wikimedia.org/T290447

also has the capability to refute and mitigate stereotypes as well as provide insights into human behavior. In an ongoing project examining tweets about Black Lives Matter (in revision for PNAS [7]), we use NLP models to reveal the prominence of positive emotions like hope and optimism, offering evidence to refute stereotypes of protesters as exclusively perpetuating anger and outrage. Future projects can develop methods for rewriting text to reduce stereotyping and models for characterizing how people are described along other affective dimensions.

Fairness and discrimination in NLP systems

While the ability of machine learning models to recognize patterns that are difficult for humans to detect opens avenues for research in prejudice and manipulation, it also can result in direct harm when deployed AI systems absorb stereotypes and historical injustice. Understanding what tangible harms can result when NLP models absorb prejudice and stereotypes requires engaging with practitioners who deploy systems and people who use them or are affected by them. My work towards understanding and preventing potential harms from learned biases has included surveying how NLP literature has engaged with race and racism as well as investigating the risks and benefits of deploying NLP models in a high-stakes setting: child welfare cases.

Investigating racial bias in NLP models Gender bias has become a well-studied topic in NLP, but substantially less work examines race. Our survey of ACL literature highlights examples of how racial biases manifest at all stages of NLP pipelines and identifies limitations of current work [9]. NLP research has only examined race in a narrow range of tasks with limited or no social context and failed to engage with people traditionally underrepresented in STEM and academia.

This survey identifies numerous areas for future work, including incorporating social context and engaging with people involved in NLP pipelines in order to understand the societal effects of NLP on marginalized populations and prevent harms. For the past year, I have been acting on some of the insights developed in this survey by collaborating with the Allegheny County Department of Human Services to investigate how the deployment of NLP technology could impact the services they offer to community members. There is intense interest in using NLP technology in child welfare settings, which often involve large amounts of text that are too numerous for overworked caseworkers to review by hand. NLP tools like information extraction, summarization, and named entity recognition can aid caseworkers in quickly identifying relevant information, while models trained to predict specific outcomes can help identify possible sources of risk and support. However, models trained on human-generated text and decisions are liable to absorbing and amplifying human prejudice and can exhibit systemic bias, such as performance gaps for people with different demographic characteristics. My initial work in this area has involved investigating how incorporating text data into existing predictive models impacts model fairness, using metrics like calibration and accuracy equity, in order to uncover possible prejudices in the text. While our current work is focused on child welfare, this research is also applicable to other domains involving expert-written notes and high-stakes decisions, such as healthcare [31]. I intend to continue this collaboration and further investigate the potential benefits and harms of implementing NLP technologies in child welfare settings, as well as expand this work to other related applications.

Throughout my PhD, I have developed multi-disciplinary collaborations with researchers from diverse backgrounds as well as people outside of academia in order to pursue impactful research directions. I am excited about expanding the capabilities of modern NLP to accomplish complex real-world tasks while minimizing potential harms.

References

[1] **Anjalie Field**, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. Framing and agenda-setting in Russian news: A computational analysis of intricate political strategies. In *Proc. of EMNLP*, pages 3570–3580, 2018.

[2] **Anjalie Field** and Yulia Tsvetkov. Unsupervised discovery of implicit gender bias. In *Proc. of EMNLP*, pages 596–608, 2020.

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

[4] **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.

[5] 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.

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

[7] **Anjalie Field***, Antonio Theophilo*, Chan Young Park*, Jamelle Watson-Daniels, and Yulia Tsvetkov. Emotion analysis and the role of positivity in #BlackLivesMatter tweets. *Working Paper*, 2021.

[8] Mengzhou Xia, **Anjalie Field**, and Yulia Tsvetkov. Demoting racial bias in hate speech detection. In *Proc. of Workshop on Natural Language Processing for Social Media at ACL*, pages 7–14, 2020.

[9] **Anjalie Field**, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. A survey of race, racism, and anti-racism in NLP. In *Proc. of ACL*, pages 1905–1925, 2021.

[10] David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. A just and comprehensive strategy for using NLP to address online abuse. In *Proc. of ACL*, pages 3658–3666, 2019.

[11] Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

[12] Sachin Kumar, Shuly Wintner, Noah A Smith, and Yulia Tsvetkov. Topics to avoid: Demoting latent confounds in text classification. In *Proc. of EMNLP*, pages 4153–4163, 2019.

[13] 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.

[14] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proc. of NeurIPS*, pages 4349–4357, 2016.

[15] Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. Powertransformer: Unsupervised controllable revision for biased language correction. In *Proc. of EMNLP*, pages 7426–7441, 2020.

[16] Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. Controlled text generation as continuous optimization with multiple constraint. In *Proc. of NeurIPS*, 2021.

[17] Kate Starbird, Ahmer Arif, and Tom Wilson. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operation. In *Proc. of CSCW*, 2021.

[18] Maxwell McCombs. The agenda-setting role of the mass media in the shaping of public opinion. In *Proceedings of the 2002 Conference of Mass Media Economics, London School of Economics*, 2002.

[19] Robert M Entman. Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173, 2007.

[20] Aman Tyagi*, **Anjalie Field***, Priyank Lathwal, Yulia Tsvetkov, and Kathleen M Carley. A computational analysis of polarization on Indian and Pakistani social media. In *Proc. of SocInfo*, pages 364–379. Springer, 2020.

[21] Arunesh Mathur, Angelina Wang, Carsten Schwemmer, Maia Hamin, Brandon M. Stewart, and Arvind Narayanan. Manipulative tactics are the norm in political emails: Evidence from 100k emails from the 2020 u.s. election cycle. *Working Paper*, 2021.

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

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

[24] 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.

[25] Natalie Schluter. The glass ceiling in NLP. In *Proc. of EMNLP*, pages 2793–2798, 2018.

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

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

[28] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *Proc. of NAACL*, pages 8–14, 2018.

[29] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proc. of NAACL*, pages 15–20, 2018.

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

[31] Nupoor Gandhi, **Anjalie Field**, and Yulia Tsvetkov. Improving span representation for domain-adapted coreference resolution. In *Proc. of Workshop on Computational Models of Reference, Anaphora and Coreference at EMNLP*, pages 7–14, 2021.