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. In contrast, my work centers the role of people in modern NLP by developing computational approaches for identifying 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 challenges, 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. Content like hate speech and misinformation can cause physical and societal harms that are liable to be amplified by NLP models trained on online data. Much work in NLP has aimed to address these concerns through supervised classification tasks, but approaches relying on annotated data fail to detect content that lacks formal definitions, is difficult for annotators to recognize, or requires modeling deeper pragmatics, like the author's intent. 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.

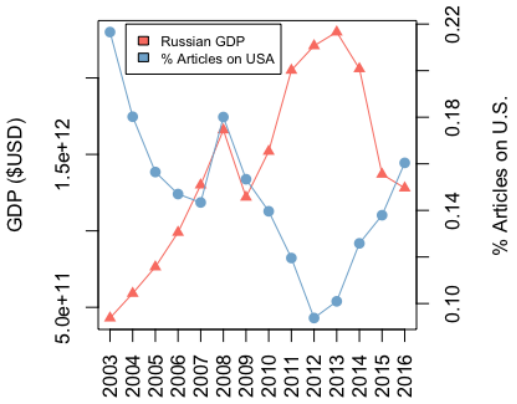

Figure 1: We uncover propaganda by showing Russian news coverage of the U.S. is strongly negatively correlated with the Russian economy [1].

Unsupervised models for detecting bias My work uncovers veiled gender bias through an unsupervised approach 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 the addressee and examining predictive features [2]. Deep learning has the capacity to process large-scale data 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 [10] and also incorporates adversarial training for demoting latent confounds [11] to guide the model to learn pragmatic features predictive of bias. There are numerous directions for future work, including further integration of causal inference and deep learning, improved methods for demoting confounds, and identification of distant supervision tasks. Furthermore, as subtle manifestations of bias are often unintentional, future research could focus on generating less harmful rephrasings.

Characterizing manipulation strategies in multilingual text While harms in text can be un-

intentional results of implicit bias, authors can also intentionally write biased content to manipulate readers, such as Russian-government-affiliated accounts disingenuously tweeting about U.S. elections [12]. Common NLP approaches like fact-checking fail to detect more subtle strategies, e.g. propagating content that is truthful but also polarizing or distracting. In contrast, 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 Russian newspaper articles using Granger causality to target agenda-setting and cross-lingual lexicon projection to target framing [1, 13, 14]. 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 examining polarization on Twitter in India following a terrorist attack [15]. Both projects involved non-English data and open-ended research questions that were difficult to address through data annotations. Instead, I developed distantly-supervised methods for categorizing and comparing text with external events.

While this work sheds new light on manipulation strategies, future research is needed to understand their effects on readers, which we can estimate through user studies and analyses of corresponding data sources, including social media and surveys. These directions require developing new NLP technology that can, for example, model differences between social and mainstream media, estimate information spread, and jointly process text with GIFs, memes, and other visual data. Further, collection efforts have resulted in much available data from manipulation campaigns that NLP analysis could derive insights from, such as 2020 U.S. election campaign emails [16]. Finally, with the growth of NLP models that generate highly fluent text more research is needed to understand what harms models absorb from disingenuous training data and how they can be mitigated.

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

Text describing people is liable to perpetuating bias, stereotypes, and prejudice [17, 18], which can cause harm, especially when amplified by NLP models [19, 20]. Analyzing these phenomena can lead to fairer NLP and has many applications. In text like news, encyclopedias, or performance reviews, authors often strive for objectivity, but implicit biases can result in unintentional 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: Alan Turing's Wikipedia page in different languages. *accepted* in the English edition 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).

Directly examining how people are described requires developing computational models capable of capturing subtle connotations aligned with social theories about affect and stereotypes. My approach develops models to score people portrayals along dimensions of power, agency, and sentiment, which behavorial science studies have identified as the most important axes of affective meaning [21]. Methodology includes integrating off-the-shelf word-level annotations with pre-trained contextualized embeddings to leverage both prior work on developing annotated word lexicons and the abilities of modern NLP to capture context [4]. Alternative methods include projecting entity embeddings into constructed affective subspaces [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: verbs 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 led 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 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 for mitigating stereotypes can adapt methods using annotated data or lexicons to new domains [22, 23].

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. 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 has examined race. Our ACL Anthology survey highlights examples of how racial biases manifest at all stages of NLP model 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. Thus, there are 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 their child welfare services. There is intense interest in using NLP in child welfare settings, which often involve too much text for overworked caseworkers to review by hand. NLP tools like information extraction and summarization 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 [24]. 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

[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

to other related applications.

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*, 2018.

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

[3] **Anjalie Field** and Yulia Tsvetkov. Entity-centric contextual affective analysis. In *Proc. of ACL*, 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*, 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*, 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*, 2021.

[10] 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.

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

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

[13] 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.

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

[15] 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*. Springer, 2020.

[16] 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.

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

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

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

[20] 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*, 2018.

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

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

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

[24] 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*, 2021.