Toward Responsible, Inclusive, and Robust Natural Language Processing

JIEYU ZHAO, DEPARTMENT OF COMPUTER SCIENCE, UCLA

Over the past few years, deep neural networks have reformed artificial intelligence. Nowadays, machine learning (ML) models can even outperform humans in some natural language processing (NLP) tasks, such as machine comprehension. However, despite the wide usage and great success, an NLP system can capture spurious features from the data it is trained on and would not generalize well. Such a problem will become much more severe in some high-stakes cases, as the system can become reliant on undesired sensitive features, resulting in unfair decision making. For example, an automatic resume filtering system might implicitly select candidates based on gender or race, amplifying existing social disparities [1, 16]. These limitations in advanced NLP effectively hinder access to people from specific demographics, namely those of minority backgrounds. As someone from underrepresented groups, my long-term research goal is to build accountable and responsible NLP techniques that are available to people from diverse backgrounds. Specifically, I plan to 1) develop a better understanding and interpretation of model behaviors, 2) contribute knowledge that guides or intervenes in existing models, and 3) apply NLP techniques in applications that foster real social good applications.

1 Prior Research Achievements

My vision is supported by my prior research in analyzing and mitigating societal biases in NLP models and applications. In the past, I have contributed to developing new models and algorithms, and injecting domain knowledge to intervene in model learning and inference. My research pioneered the direction of social equity-based NLP, fostering a promising research field that caught on quickly throughout the community. Back in 2017, there was only one paper about social equity at ACL. Now, there are several individual conference tracks or workshops on the topic of fairness in NLP. I have published papers in top NLP venues, such as EMNLP, ACL, and NAACL and some have been incorporated into machine learning libraries such as AllenNLP models and HuggingFace [1], which can be easily accessible by other researchers. My paper won the **Best Paper Award** at EMNLP 2017, and was covered in the media, including by *WIRED* and *The Daily Mail*. Collectively, my papers have been cited more than 1,600 times since 2017, according to Google Scholar; some have been included in the textbook *Speech and Language Processing* [7] and taught in NLP and Ethics courses in CMU and Stanford [20, 6]. My work has also been widely cited outside NLP, such as in computer vision [23], sociology [3], and in biomedical literature [22].

1.1 Model bias detection

With an increasingly large number of NLP applications having an impact on our daily life, it is essential to guarantee that those models serve all users and do not make decisions discriminating against a particular demographic group. Such intention first raises the question: How can we effectively detect biases in a model? To answer this question, my work focuses on: 1) providing new ways to quantify biases, 2) revealing bias amplification in existing models, and 3) understanding biases in cross-lingual scenarios.

New bias quantification. My research stands as one of the pioneering works to understand the societal biases in NLP: My colleagues and I are among the first to define fairness in NLP. Previously, one of the biggest challenges in the community was the absence of good ways to quantify bias. For this, I have proposed LOGAN, a new bias detection technique based on clustering [12]. I have also contributed the first few datasets for bias quantification: Based on counterfactual method, I released WINOBIAS for bias detection in coreference resolution [17]. The work has been widely cited, inspiring new dataset curation [21] as well

[1] They have become basic building blocks of modern NLP models and are widely used in academia and industry.

as new research on detecting biases in widely used tasks such as in machine translation [19]. Besides, I developed datasets such as, MULTILINGUAL BIOSBIAS for text classification [14] and WIKIGENDERBIAS for neural relation extraction [2]. All these efforts filled the gap of lacking convenient ways to evaluate biases, which in return spurred the development of better NLP models.

Bias amplification. It has become a common belief that models run the risk of discovering and exploiting societal biases present in the underlying corpora [1]. But to our surprise, a model not only duplicates such biases, but further magnifies them. Without properly quantifying and reducing such magnification, a broad adoption of these models can significantly impact certain users. In one of my first-authored papers [16], I find that (a) datasets for vision-and-language tasks contain significant gender bias and that (b) models trained on these datasets further **amplify** existing bias. This work is the *first* one to reveal social bias amplification in NLP (and in ML more broadly) and has been widely discussed ever since. It has been reviewed by online media in both English and Chinese. Now the bias amplification issue is widely recognized by the community and has stimulated several follow-up works in different research areas [4, 10]. This paper won the *Best Paper Award* at EMNLP 2017 and has been cited more than 500 times.

Cross-lingual scenario. One of my research goals is to make NLP models available to people from different communities. In terms of the societal bias issue, most of the existing literature is conducted in English, the dominant language in NLP. However, NLP techniques must be inclusive and applicable to more than one language. One line of my research conducted bias analysis under multilingual cases [14, 24]. The results show that despite linguistic differences (e.g., grammatical gender in Spanish), biases commonly exist in different languages and exhibit stereotypical behavior similar to that of English. In addition, many existing cross-lingual transfer learning tasks heavily rely on multilingual embeddings, and in [14], I demonstrated that biases in multilingual embeddings affect cross-lingual transfer learning tasks.

1.2 Model bias mitigation

My research has shown that societal biases in NLP models can arise from various sources such as implicit bias in the training data [17], (contextualized) word representations [15], and models trained on biased datasets can further amplify the bias [16]. Based on those discoveries, I have been able to innovatively adapt several methodologies from other domains to the bias mitigation scenario. Those methods touch different layers of an NLP model and can be categorized as follows:

Modifying training resources. Modern NLP models often rely on a collected training dataset and learn the underlying representations of input examples. Unfortunately, the training dataset is usually imbalanced and the representations learned also implicitly inherit the biases, both of which eventually contribute to model biases. Such insights motivate me to reduce model biases from both the *data* and *representation* perspectives. For the former, I proposed to generate an auxiliary training dataset based on some rules [17, 15]. It eliminates bias without significantly affecting the overall model performance. For the latter, I came up with the first method to learn word embeddings with protected attributes by disentangling them from other information [17]. For contextualized vectors, I show that they encode and propagate gender information unequally. A lightweight method by balancing the context for such contextualized embeddings regarding gender can help to reduce the bias in the test time without the need to retrain them [15]. Those techniques have inspired many follow-ups, from new bias mitigation in embeddings [8] to in new domains [18].

Adding inference constraints. In many cases, we can only access an already trained model without knowing the underlying training procedure, which makes bias control by modifying the training procedure impossible. I have proposed to inject corpus-level constraints for calibrating model predictions to reduce bias amplification. In [16, 5], my colleagues and I design algorithms based on Lagrangian relaxation and posterior regularization, respectively, for the inference. Those techniques are widely used in many tasks such as

in semisupervised learning, but we are the first to adopt them for bias mitigation. Such techniques do not require model retraining but decrease the magnitude of bias amplification by a large magnitude (e.g., almost no bias amplification left after posterior regularization) with negligible performance loss on the original task. I also created tutorials for these methods so that they can be better understood by and assist others.

Injecting ethical interventions. Humans can learn and adjust their behavior after learning from new instructions. Inspired by this, McCarthy envisioned a machine that can take declarative knowledge as input and make decisions based on that [11]. But such a great vision remains elusive due to various challenges in that era. Now, with the advancement in NLP, I revisit this hypothesis and explore the idea of guiding a model by instruction. To be specific, I design the Linguistic Ethical Interventions task, where the goal is to amend a model's unethical behavior by communicating context-specific principles of ethics and equity to it [13]. This work stands as one of the pioneer efforts to formalize and study the effectiveness of natural language interventions to amend model behavior. And I present the task as a challenge for our community.

2 Ongoing and Future Research

My research builds computational methodologies to characterize societal biases in modern NLP models as a means of discovering biased model behaviors and proposing methods to mitigate those biases. I have come to deeply understand the common existence of biases and the difficulty of addressing them [9]. The former makes the topic an interdisciplinary subject that requires a general and trustworthy way to detect the bias, while the latter calls for a robust and accountable way to do the mitigation. I have been fortunate to collaborate with people from different backgrounds, covering NLP, computer vision, data mining, and information retrieval, which provides a basis for the development of my future work. In short, my research agenda commits to designing a general framework to advance accountable and ethical NLP models, which both benefits advanced methodologies in computer science and addresses essential interdisciplinary problems. I am taking three directions which extend my expertise and move beyond the existing literature.

1. Understanding model behaviors and conducting early detection of model biases. There have been great efforts to improve existing models and a lot of new models are being proposed. With such a high volume of new NLP models each year, it is of great importance to understand the underlying reasoning and logic of model behavior, which can provide much more transparent and accountable insights into each model. While there exists a line of research on detecting and mitigating bias in different NLP models, understanding precisely what leads to these biases is underexplored: this is the realm of model interpretability. Such interpretation tools will not only tell us why there is bias in the model, which elicits an early detection of the bias, so that we do not need to wait for the model outputs for bias detection, but also provide a method for early intervention. For example, if a resume filtering model uses sensitive attributes (e.g., gender) for a hiring decision, we can adjust it by adding constraints to its inference procedure.

2. Fusing additional knowledge to boost model accountability. While we are entering the era of automated decision making, human decision makers are typically still in control, especially in high-stakes settings, as the models often lack all relevant knowledge, and a pure model scaling may be insufficient. We need new models with the ability to understand "commonsense" that would come in different formats, such as in an instruction document style or in a more interactive approach. I plan to continue this line of research by finding more efficient ways to control the model, such as by using advanced trigger techniques. Another way to intervene in models is to keep human experts involved in decision-making loops; in this way, we can also provide more specific knowledge for the models. I plan to build interactive systems that can take human feedback and update the models based on the feedback, mitigating bias in the system, and explicitly flagging it for human decision makers. This direction interacts with direction (1) and combines to form a more comprehensive solution, building on my expertise and exploring new directions.

3. Building NLP for social good. In the long term, I strongly believe that advanced NLP or machine learning techniques can play a vital role in the real world; a significant case is to adapt those models to benefit the social good. While such advanced techniques have had a major impact in our daily life – for example, we can access an enormous set of resources by simply using a mobile device – such resources are unavailable to many people, such as those with disabilities. For these populations, those seemingly-easily accessible devices can become substantial barriers – during Covid-19, when everything was moved online, people with a visual disability could hardly get anything done. In the future, I am eager to explore developing NLP techniques for social good and applying my expertise in other real-world applications, such as helping people with disabilities. I want to collaborate with people from NGOs (e.g., ADD International), to understand the real needs of people from neglected groups and provide technical support.

The methods I am developing can be applied to build accountable NLP models beyond social bias perspective. These directions lie in the intersection of ML, NLP, and other disciplines such as the humanities and sociology. I look forward to exploring these directions and expanding my collaborations!

[1] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS*, 2016.

[2] A. Gaut, T. Sun, S. Tang, Y. Huang, J. Qian, M. ElSherief, **J. Zhao**, D. Mirza, E. Belding, K.-W. Chang, et al. Towards understanding gender bias in relation extraction. In *ACL*, 2020.

[3] A. Hagerty and I. Rubinov. Global ai ethics: a review of the social impacts and ethical implications of artificial intelligence. *arXiv preprint arXiv:1907.07892*, 2019.

[4] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018.

[5] S. Jia, T. Meng, **J. Zhao**, and K.-W. Chang. Mitigating gender bias amplification in distribution by posterior regularization. In *ACL*, 2020.

[6] D. Jurafsky. Ethical and Social Issues in Natural Language Processing. https://web.stanford.edu/class/cs384/.

[7] D. Jurafsky and J. H. Martin. Speech and Language Processing. https://web.stanford.edu/~jurafsky/slp3/.

[8] M. Kaneko and D. Bollegala. Gender-preserving debiasing for pre-trained word embeddings. In *ACL*, 2019.

[9] F. Khani and P. Liang. Removing spurious features can hurt accuracy and affect groups disproportionately. In *FAccT*, 2021.

[10] K. Leino, M. Fredrikson, E. Black, S. Sen, and A. Datta. Feature-wise bias amplification. In *ICLR*, 2019.

[11] J. McCarthy et al. *Programs with common sense*. RLE and MIT computation center, 1960.

[12] **J. Zhao** and K.-W. Chang. Logan: Local group bias detection by clustering. In *EMNLP*, 2020.

[13] **J. Zhao**, D. Khashabi, T. Khot, A. Sabharwal, and K.-W. Chang. Ethical-advice taker: Do language models understand natural language interventions? *ACL Findings*, 2021.

[14] **J. Zhao**, S. Mukherjee, S. Hosseini, K.-W. Chang, and A. H. Awadallah. Gender bias in multilingual embeddings and cross-lingual transfer. In *ACL*, 2020.

[15] **J. Zhao**, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang. Gender bias in contextualized word embeddings. In *NAACL (short)*, 2019.

[16] **J. Zhao**, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017.

[17] **J. Zhao**, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *NAACL*, 2018.

[18] J. H. Park, J. Shin, and P. Fung. Reducing gender bias in abusive language detection. In *EMNLP*, 2018.

[19] G. Stanovsky, N. A. Smith, and L. Zettlemoyer. Evaluating gender bias in machine translation. In *ACL*, 2019.

[20] Y. Tsvetkov and A. W. Black. Computational Ethics for NLP. http://demo.clab.cs.cmu.edu/ethical_nlp2020/.

[21] K. Webster, M. Recasens, V. Axelrod, and J. Baldridge. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *TACL*, 2018.

[22] Y. Xie, M. Chen, D. Kao, G. Gao, and X. Chen. Chexplain: Enabling physicians to explore and understand data-driven, ai-enabled medical imaging analysis. In *CHI*, 2020.

[23] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019.

[24] P. Zhou, W. Shi, **J. Zhao**, K.-H. Huang, M. Chen, R. Cotterell, and K.-W. Chang. Examining gender bias in languages with grammatical gender. In *EMNLP*, 2019.