My research aims to make **equitable natural language processing (NLP) systems adaptable to individuals and groups**. Despite the ubiquity and popularity of large language models (LLMs), numerous studies have shown that their benefits are not equal across cultures and individuals. This is partly because today's LLMs are primarily optimized to generate the most average responses for the most average person based on training data that is insufficient to learn nuanced, inclusive responses. For example, when responding to questions like 'What precautions should I take when visiting Saudi Arabia?', language models should consider the questioner's cultural background, age, gender, sexuality, and personal preferences, all factors that significantly influence the kinds of situations they may encounter.

I address these challenges by (1) **centering the sociocultural context in modern NLP** and developing computational approaches that **identify fairness issues** that arise from failures to incorporate sociocultural factors and (2) **developing novel socially aware NLP models** that explicitly incorporate social and cultural contexts around language. Specifically, my research addresses the following key areas at the intersection of AI Ethics, NLP, and computational social science (CSS):

- **Measuring social biases in sociotechnical system (AI Ethics)**. How do we identify and robustly measure social biases in language technologies, at *all* stages of development? [8, 29, 10]
- **Incorporating social contexts in NLP models (NLP)**. How do we develop socially aware NLP models that incorporate semantic and pragmatic sociocultural knowledge? [27, 21, 32, 28, 26, 17]
- **Explaining social phenomena with NLP (CSS)**. How can we use socioculturally aware NLP models to better understand people and cultures through the lens of language? [30, 31, 3]

In addressing these questions, I develop general and theoretically grounded computational methods, model architectures, analysis algorithms, and datasets. Further, I employ advanced ML/NLP techniques to showcase the practical applications of language technologies. My work explores **a variety of languages with cross-lingual methods** and multilingual models, spanning structured sources like news articles and Wikipedia to the unstructured landscape of social media. **Bridging social science and language technologies**, my research sheds new light on our understanding of people and cultures and develops state-of-the-art technologies to equitably serve diverse users.

Robustly Measuring Social Biases in Language Technologies

It is well established that NLP models learn and amplify social biases [2, 13]. While considerable work addresses social biases in language technologies, it generally focuses on a limited set of biases (e.g., gender or racial bias) [9] within limited scenarios [23] in a single language (primarily, English) [15]. This makes findings less generalizable and less robust [11]. I aim to develop novel computational methods to identify and measure more diverse social biases more robustly and comprehensively.

Controlled Multilingual Affect Analyses for Measuring Social Biases on Wikipedia As a step towards developing robust algorithms to analyze social biases more holistically, I introduced a series of methods and studies that define and quantify various dimensions of social biases, including biases towards LGBT, non-binary, and intersectional identity groups in Wikipedia biographies across multiple languages [29, 10]. My collaborators and I proposed **biography matching algorithms grounded in causal inference methods** to control for confounding factors [10] and a **multilingual affective analysis model** [29] that leverages crosslingual contextual sentence embeddings to measure implied affect towards a person along dimensions of sentiment, power, and agency. We built the **first multilingual dataset for the contextual affective analysis task** to train the model. Our analysis reveals that Russian articles tend to use verbs with more negative connotations when describing LGBT people than English or Spanish articles, confirming different perceptions across cultures.

The Wikimedia Foundation, the primary stakeholder with power to use our research to mitigate social biases in narratives about people, recognized our work with the **Wikimedia Foundation Research Award of**

the Year for 2023 and is implementing our methods.[1] Our work has also received interest from the **journalism community**, leading to a collaboration with *The Washington Post* to examine anti-Black discrimination on Chinese social media [3].

Tracking the Trails of Social Biases Leading to Unfair NLP Models Another important direction towards more comprehensive understanding of social biases in NLP systems is to investigate the end-to-end process of model development rather than individually assessing each stage [11]. Recently, we studied how political biases in training data, yet another less-studied challenge prevalent in NLP systems, propagates to LLMs and eventually leads to unfair NLP models [8]. We devised a framework, grounded in established *political compass tests* from political science to **measure political leanings of LMs** by leveraging a stance detector to assess the models' positions on various topics. Subsequently, we created training datasets with controlled political leanings and further pretrained models separately to reveal that political biases in training data indeed propagate to LMs. We then finetuned the models with varying political beliefs on the same downstream task datasets to investigate if LLMs' political biases impact their downstream behavior and fairness.

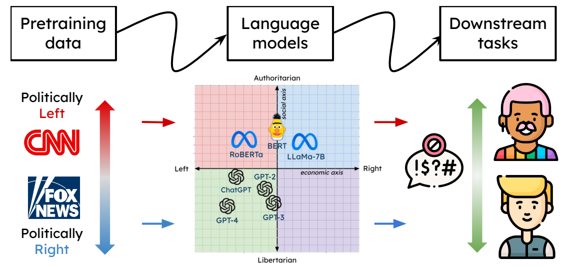

Figure 1: We measure political biases in LLMs and systematically examine how they lead to unfair NLP models for downstream tasks.

We found that **LLMs have distinctly different political leanings, and their biases consequently affect the kinds of content that models classify** as hate speech and misinformation. For example, models with left-leaning biases showed increased sensitivity to hate speech targeting minority groups such as Women and LGBTQ+ individuals. Conversely, models with right-leaning biases exhibited greater sensitivity to hate speech directed at white Christian men. Our work was awarded the **Best Paper Award at ACL 2023** and received extensive media attention, including coverage and interviews by *The Washington Post* [34] and the *MIT Tech Review* [12].

Future Work. My research identified major open questions in modeling human-centered NLP, especially focusing on cultural underpinnings of human biases and values. For example, I am currently investigating *cultural biases* in LLMs, another under-explored area of social bias in NLP. Although recent studies have explored which culture LLMs are most aligned with based on their values [4, 1], there is no work quantitatively measures how well LLMs adapt to and serve different cultures. Measuring cultural biases of LLMs requires defining and formulating cultural knowledge that is operationalizable by NLP methods, building a challenging benchmark dataset that can test reasoning ability for both explicit cultural knowledge and implicit cultural norms, and designing metrics that can capture both absolute cultural understanding and relative equity across cultures.

Towards robust measurement of social biases, I intend to continue developing methodologies for controlled and systematic analysis. As one example, we are investigating how different cultures and their biases influence content selection in Wikipedia articles. Our primary focus is in developing NLP methods to automatically identify information gaps between multilingual texts. Analyzing these gaps will provide valuable insights into how content selection can reflect cultural biases and how that, in turn, affects portrayal disparities among articles in different languages.

Incorporating Social Contexts to Improve NLP Models

For machine learning (ML) systems to be equitable and adaptable to individuals, they need a way to represent and account for social contexts [13]. Here, the key challenge is that much social information is not explicitly

[1] https://phabricator.wikimedia.org/T290447

stated in language. As a result, social context is missing in most training datasets and in current models [14]. My work focuses on (1) building **datasets that encode social contexts** along with the texts, including information about writers and readers and the social settings in which the texts are contextualized, and (2) developing **new ML models that integrate** such information.

Community Context for Norm Violation Detection Today's automated tools for moderating online communities (e.g., hate speech detectors) do not take social context into account [22]. I hypothesized that incorporating explicit knowledge about a community and its rules is crucial for detecting community norm violations more accurately. To validate this hypothesis, we **collected NormVio** [27], a dataset that contains 52K comments from Reddit, their communities (i.e., subreddits), their respective community rules, prior conversation information, and labels indicating whether they violated any community rules and were moderated by human moderators. We then introduced **context-sensitive norm violation classifiers**; unlike existing hate speech classifiers that rely solely on text input, they consider community-specific information. Our best model outperformed context-insensitive baselines in detecting norm violations by nearly 50%, and our models can pinpoint specific violated rules in a community. Context-sensitive classifiers thus provide a key **practical assistive technology**: they help human moderators identify inappropriate content for their specific communities and better communicate their rationale to users, lessening the burden of managing the overwhelming influx of new posts and comments. This work **led to a startup's interest** in developing a similar model for other platforms that suffer from intractable amounts of toxic comments, e.g., real-time chat platforms like Twitch, resulting in a collaboration paper accepted at EMNLP 2023 [21].

Generative Zero-Shot Classifier with Text Labels for Personalization Since a limited number of datasets offer social context, zero-shot classifiers that can account for social context without training data can be especially valuable. We introduced a *generative classifier* for zero-shot classification [17] that enables the simple personalization and adaptation of models by incorporating social context through a text label (Figure 2). For example, a comment "go get it girl" might be empowering when addressed to a woman but sarcasm when addressed to a man. Our model calculates the probability of generating the comment, given the contextual text label, such as "The comment written by a woman empowers the addressed woman", to determine whether the comment is empowering. Our model, evaluated across 18 different tasks including hate speech and empowerment prediction [32], shows a better classification performance than strong in-context learning baselines. For this line of work – making LLMs more socioculturally aware – I received CMU's competitive **K&L Gates Presidential Fellowship for Ethics and Computational Technologies**.

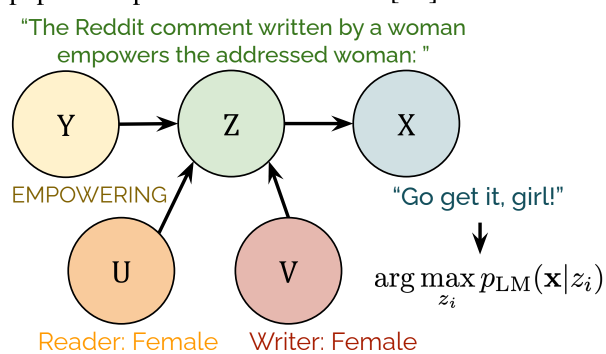

Figure 2: Our generative framework measures the LLMs' likelihood of generating input text $\mathbf{x}$, conditioned on natural language descriptions of labels z to incorporate social contexts into LLMs.

Future Work. My work has advanced the development of socioculturally aware NLP models to make them more effective and equitable, but much work remains to be done. In addition to classification, I am particularly interested in designing LLMs that can *generate* responses that are more appropriate and useful by considering users' sociocultural backgrounds. I am currently working on contextualizing reinforcement learning from human feedback (RLHF), known to be a critical step in aligning models with human preferences, with social contexts. Specifically, we constructed a set of paired Reddit comments with human preference labels indicated by users' upvotes across diverse subreddits. We are testing whether contextualizing models during RLHF helps the model generate responses that are better suited to each subreddit.

Leveraging NLP Models to Explain Social Phenomena

People communicate through language, and social scientists analyze language to better understand society. Despite the remarkable advances in NLP, not all of them have been adopted by social scientists largely because many language technologies are not developed with real-world applications in mind [20]; as such, they might not perform adequately for use in a new target domain [16] and might overlook essential requirements for social scientists, such as interpretability [24]. Bridging the gap between NLP and computational social science, my research aims to **identify shortcomings** in NLP models for CSS [31], propose **NLP solutions** to close the gap [30], and showcase how to **use state-of-the-art NLP** methods to gain new knowledge about society [30, 31].

Analyzing the Driving Forces of Activism using Robust Emotion Classifiers In my 2022 **PNAS** paper [30], we showed how to analyze the relationship between social movements and emotions expressed on social media using domain adaptation on existing emotion classification data and models. Specifically, we compared various unsupervised and few-shot training methods for domain adaptation to provide **guidelines for social scientists** who want to use NLP models on their own target tasks and domains. With our domain-adapted emotion classifiers, we collected and analyzed 34M tweets posted during the 2020 Black Lives Matter (BLM) protests in collaboration with the Data for Black Lives organization. We found that expressions of positivity, like hope and camaraderie, influence the movement more than negative ones, like anger (see Figure 3), thereby **countering a common stereotype of "angry Black people."**

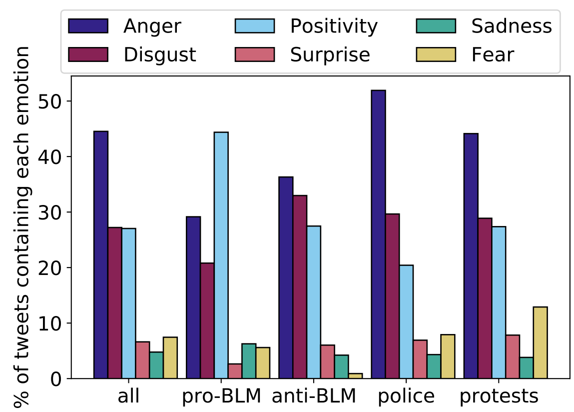

Figure 3: Emotion distribution of tweets by hashtags reveals that positivity is the main emotion pro-BLM tweets exhibit in contrast to other hashtags.

Challenges of NLP models in Detecting Information Manipulation In another work [31], we analyzed challenges and opportunities of NLP models used to detect information manipulation by examining Russian media during the 2022 Ukraine-Russian war. We **collected a dataset** of 10M+ Russian social media posts by state-affiliated and independent media outlets along with public reactions to them (e.g., likes and comments), and we applied state-of-the-art NLP models, such as topic modeling and media framing classifiers. While we identified numerous opportunities for NLP research to make positive contributions in combating real-world information manipulation campaigns (e.g., **uncovering agenda setting and framing strategies** of stat-affiliated media), we also recognized challenges in developing more deployable and interpretable technology for use in evolving situations. This project generated a high demand for our dataset (**20+ requests**), underscoring its value in advancing research on real-world misinformation and information campaigns. One data request resulted in a collaborative effort and **funding from DoD** ($160K) aimed at using language technologies in identifying evolving narratives in information operations.

Future Work Another important factor that hinders the wide adoption of NLP methods is their interface [35]; for example, social scientists need to figure out how to finetune and run the models. However, with the introduction of generative LLMs such as ChatGPT, language technologies now have an exciting potential to provide powerful reasoning ability in convenient, easily learnable ways [36]. I plan to continue investigating and demonstrating various ways to use LLMs to address social science research questions. As one example, collaborators and I are now working on developing computational approaches to identify social norms that govern language in online communities. We are actively exploring the utility of LLMs in interpreting a community's recognition signals (e.g., the number of upvotes on Reddit) to decipher which language is more/less appreciated by the community.

Future Directions

Aligned with my long-term goal of building socioculturally aware NLP models to make them equitable and accessible, I plan to expand my research in various interdisciplinary directions, including the following.

AI Safety and Public Policy My research has shown how social biases in NLP models can disproportionally impact users [29, 17]. My primary goal in investigating such biases is to prevent harm, to individuals and larger social entities. I aim to contribute to this crucial mission by focusing on AI Safety, developing rigorous evaluation methods and benchmarks. One of my primary objectives is ensuring these evaluation methods get widely adopted, not just within the research community but also in the industry. I intend to make them comprehensive and applicable to a broad range of domains and problems. This will help ensure the safety and trustworthiness of models that millions of people use on a daily basis. In line with my commitment to practical impact, I will also explore the concept of *active evaluation* and develop computational strategies for updating evaluation methods and data to adapt to changes in models and language. This approach will ensure the ongoing effectiveness of benchmarks. Finally, to ensure that advancements in AI safety research translate into real-world safety, I will collaborate with researchers and practitioners in **public policy** to explore how the measures developed within the NLP community can be effectively implemented to guide one of the most powerful yet opaque technologies ever created.

Socially Aware Multilingual Models Numerous studies have highlighted performance discrepancies and social biases exhibited by LLMs across languages [18, 6]. I intend to develop computational approaches to mitigate these performance gaps within multilingual models. One promising avenue of exploration involves implementing a *teacher-student model framework* [5] *between resource-rich and low-resource languages*, with the help of translation models. However, one foreseeable challenge in pursuing this direction is in determining what teachers can teach to students [7, 19]. For example, the answer to a question like "What is 1+1?" may be universally transferable to all languages, whereas answers to questions like "How much should I tip at a restaurant?" can vary significantly based on language and culture. To address this, I will leverage my computational expertise to collaborate with experts in **linguistics, social psychology, and anthropology** in developing more equitable and socially aware multilingual models.

Personalization and User Privacy To deploy models that incorporate users' sociocultural context in the real-world, an understanding of users' privacy requirements is indispensable [33]. Similar to personalized ads, to make this technology inviting, users should be able to control what information models can access and know why models make certain decisions [25]. In future work, I will focus on ways to build models that are controllable and interpretable by users, making them socially aware without being intrusive. Furthermore, I plan to collaborate with **Human Computer Interaction** (HCI) experts to investigate how much personalization is appropriate for users and how it should be implemented in applications.

As a faculty member, I am excited to take further steps in building equitable, inclusive, ethical, and trustworthy NLP systems and to foster multidisciplinary collaborations.

References

[1] A. Arora, L.-a. Kaffee, and I. Augenstein. Probing pre-trained language models for cross-cultural differences in values. In S. Dev, V. Prabhakaran, D. Adelani, D. Hovy, and L. Benotti, editors, *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.c3nlp-1.12. URL `https://aclanthology.org/2023.c3nlp-1.12`.

[2] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL `https://aclanthology.org/2020.acl-main.485`.

[3] S. Cahlan and J. S. Lee. Video evidence of anti-Black discrimination in china over coronavirus fears. *The Washington Post*. URL `https://www.washingtonpost.com/politics/2020/06/18/video-evidence-anti-black-discrimination-china-over-coronavirus-fears/`.

[4] Y. Cao, L. Zhou, S. Lee, L. Cabello, M. Chen, and D. Hershcovich. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In S. Dev, V. Prabhakaran, D. Adelani, D. Hovy, and L. Benotti, editors, *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.c3nlp-1.7. URL `https://aclanthology.org/2023.c3nlp-1.7`.

[5] Y. Chen, Y. Liu, Y. Cheng, and V. O. Li. A teacher-student framework for zero-resource neural machine translation. In R. Barzilay and M.-Y. Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1176. URL `https://aclanthology.org/P17-1176`.

[6] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL `https://aclanthology.org/2020.acl-main.747`.

[7] M. F. Elahi and P. Monachesi. An examination of cross-cultural similarities and differences from social media data with respect to language use. In *LREC*, pages 4080–4086, 2012.

[8] S. Feng, **Chan Young Park**, Y. Liu, and Y. Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada, July 2023. Association for Computational Linguistics. URL `https://aclanthology.org/2023.acl-long.656`.

[9] A. Field, S. L. Blodgett, Z. Waseem, and Y. Tsvetkov. A survey of race, racism, and anti-racism in NLP. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.149. URL `https://aclanthology.org/2021.acl-long.149`.

[10] A. Field, **Chan Young Park**, K. Z. Lin, and Y. Tsvetkov. Controlled analyses of social biases in Wikipedia bios. In *Proc. The ACM Web Conference '22*, 2022.

[11] S. Goldfarb-Tarrant, R. Marchant, R. Muñoz Sánchez, M. Pandya, and A. Lopez. Intrinsic bias metrics do not correlate with application bias. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.150. URL `https://aclanthology.org/2021.acl-long.150`.

[12] M. Heikkilä. Ai language models are rife with different political biases. *MIT Technology Review*. URL `https://www.technologyreview.com/2023/08/07/1077324/ai-language-models-are-rife-with-political-biases/?truid=&utm_source=the_algorithm&utm_medium=email&utm_campaign=the_algorithm.unpaid.engagement&utm_content=08-07-2023`.

[13] D. Hovy and S. Prabhumoye. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432, 2021.

[14] D. Hovy and D. Yang. The importance of modeling social factors of language: Theory and practice. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.49. URL `https://aclanthology.org/2021.naacl-main.49`.

[15] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL `https://aclanthology.org/2020.acl-main.560`.

[16] L. Kühnel and J. Fluck. We are not ready yet: limitations of state-of-the-art disease named entity recognizers. *Journal of Biomedical Semantics*, 13(1):26, 2022. doi: 10.1186/s13326-022-00280-6. URL `https://doi.org/10.1186/s13326-022-00280-6`.

[17] S. Kumar, **Chan Young Park**, and Y. Tsvetkov. Gen-z:. *arXiv preprint arXiv:2305.14326*, 2023.

[18] J. Lee, D. Lee, and S.-w. Hwang. Script, language, and labels: overcoming three discrepancies for low-resource language specialization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13004–13013, 2023.

[19] B. Y. Lin, F. F. Xu, K. Zhu, and S.-w. Hwang. Mining cross-cultural differences and similarities in social media. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 709–719, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1066. URL `https://aclanthology.org/P18-1066`.

[20] A. Macanovic. Text mining for social science–the state and the future of computational text analysis in sociology. *Social Science Research*, 108:102784, 2022.

[21] J. Moon, D.-H. Lee, H. Cho, W. Jin, **Chan Young Park**, M. Kim, J. May, J. Pujara, and S. Park. Analyzing norm violations in live-stream chat. *arXiv preprint arXiv:2305.10731*, 2023.

[22] E. Mosca, M. Wich, and G. Groh. Understanding and interpreting the impact of user context in hate speech detection. In L.-W. Ku and C.-T. Li, editors, *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 91–102, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.socialnlp-1.8. URL `https://aclanthology.org/2021.socialnlp-1.8`.

[23] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11), 2021.

[24] T. Rudas and G. Péli. *Pathways between social science and computational social science: theories, methods, and interpretations*. Springer, 2021.

[25] S. S. Sundar and S. S. Marathe. Personalization versus customization: The importance of agency, privacy, and power usage. *Human communication research*, 36(3):298–322, 2010.

[26] **Chan Young Park***, H. Ahn*, J. Sun*, and J. Seo. NLPDove at SemEval-2020 task 12: Improving offensive language detection with cross-lingual transfer. In *Proc. SemEval-2020*, 2020.

[27] **Chan Young Park**, J. Mendelsohn, K. Radhakrishnan, K. Jain, T. Kanakagiri, D. Jurgens, and Y. Tsvetkov. Detecting community norm violations in online conversations. In *Proc. EMNLP Findings'21*, 2021.

[28] **Chan Young Park***, J. Sun*, H. Ahn*, Y. Tsvetkov, and D. R. Mortensen. Ranking transfer languages with pragmatically-motivated features for multilingual sentiment analysis. In *Proc. EACL'21*, 2021.

[29] **Chan Young Park***, X. Yan*, A. Field*, and Y. Tsvetkov. Multilingual contextual affective analysis of LGBT people portrayals in wikipedia. In *Proc. ICWSM'21*, 2021.

[30] **Chan Young Park***, A. Field*, A. Theophilo*, and Y. Tsvetkov. An analysis of emotions and the prominence of positivity in #blacklivesmatter tweets. *National Academy of Sciences (PNAS)*, 2022.

[31] **Chan Young Park***, J. Mendelsohn*, A. Field*, and Y. Tsvetkov. Challenges and opportunities in information manipulation detection: An examination of wartime Russian media. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5209–5235, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.382.

[32] **Chan Young Park***, L. Njoo*, O. Stappart, M. Thielk, Y. Chu, and Y. Tsvetkov. Talkup: A novel dataset paving the way for understanding empowering language. *arXiv preprint arXiv:2305.14326*, 2023.

[33] E. Toch, Y. Wang, and L. F. Cranor. Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems. *User Modeling and User-Adapted Interaction*, 22:203–220, 2012.

[34] G. D. Vynck. Chatgpt leans liberal, research shows. *The Washington Post*. URL https://www.washingtonpost.com/technology/2023/08/16/chatgpt-ai-political-bias-research/.

[35] S. Wibberley, D. Weir, and J. Reffin. Language technology for agile social media science. In P. Lendvai and K. Zervanou, editors, *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 36–42, Sofia, Bulgaria, Aug. 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-2705.

[36] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang. Can large language models transform computational social science? *Computational linguistics*, 49(4), Dec. 2023.