Research Agenda

As a researcher I strive to **develop Natural Language Processing (NLP) technologies that foster positive social interactions and lead to more equitable systems**. My work focuses on pragmatic formalisms that measure factuality, intent, and social bias of language. For example, given the false statement "*Water boiled with garlic cures coronavirus*," my work uses commonsense reasoning to bridge the gap between what is explicitly stated and implicit knowledge like implied actions, e.g. "*people should drink garlic water*." These formalisms can support building AI systems to address critical applications like fact-checking and content moderation. I have identified three key thrusts of research to investigate:

1. **Social Commonsense Reasoning** - How well can Deep Learning models accurately capture pragmatic context like user perspective (e.g. political leanings), predict reactions or perceive emotions of people in everyday situations? [2, 20, 21]

2. **Factuality in Neural Text Generation** - How can Deep Learning models recognize text that is factually consistent with prior context and with general world knowledge? [2, 9, 12]

3. **Equity & Inclusion in Neural Language Models** - How can an Artificial Intelligence (AI) system be designed with inclusivity and interpretability in mind? Do particular neural language models lead to model behavior that might reinforce social inequities? [13, 17, 16, 22]

While essential to the design of algorithms that function well in real-world settings, these types of reasoning are not well-captured by current Deep Learning approaches. My research explains where the shortcomings lie, what are the dangers of not considering pragmatic context in machine reasoning, and how to improve reasoning capabilities using generative approaches.

Analyzing Misinformation with Social Commonsense Reasoning:

The first key aim of my research is determining how textual content impacts readers (e.g. cognitive, emotional and physical consequences resulting from a reader's interpretation of a text fragment). My goal is to use social commonsense reasoning to understand how readers perceive reliability of textual claims and what the broader societal impacts of letting claims proliferate are. This is a nuanced problem owing to the out-of-context manner in which media is usually presented [1]. Furthermore, determining factuality of claims does not reveal the likelihood of claims to spread or potential impact of claims on readers.

To analyze factors involved in risk assessment and virality of claims, I introduced a structured pragmatic formalism for interpreting implications of news headlines (Misinfo Reaction Frames) [2].

This work is grounded in prior literature on Frame Semantics [3], Theory-of-Mind [4] and machine commonsense reasoning [5]. It seeks to elucidate physical, emotional and cognitive impact of implicit messages conveyed by news headlines on readers. For the project, I developed a knowledge graph with 6 dimensions relating to impact of a news headline, including factuality, likelihood of conveyed information being spread and actions the headline may provoke a reader to take. Using international Covid-19, climate change and cancer news headlines that were verified by fact-checkers, I crowdsourced a large-scale dataset based around the defined knowledge graph for reasoning over headlines. I then used the curated dataset to develop tools for assessing unseen headlines by training transformer-based models [6, 7]. Initial findings from statistical hypothesis testing (specifically a A/B study) I conducted indicate that the approach is effective at calibrating users' trust in news headlines - **users trust misinformation headlines less** after viewing machine-generated Misinfo Reaction Frame implications.

I recently presented the work at ACL 2022, and am working with collaborators in Security & Privacy research at the University of Washington to deploy the developed tools as an browser extension application (The in-progress demo using a model trained on Misinfo Reaction Frames can be seen in Figure 1). This will empower users to think more critically about online content and share their own interpretations of implicit messages conveyed by digital media. I believe such tools for detecting false or harmful language could have wide-ranging societal benefits. One potential case study is an ongoing project I am engaged in with Cynthia Breazeal[1] and researchers from MIT Media Lab[2] where we are exploring the use of social agents to help children develop reading comprehension skills. Here factually- and ethically-constrained generation could prevent potential harms of conversational AI [8] and ensure child-appropriateness of dialogue output by educational social agents.

Evaluating Factuality of Machine-generated Text:

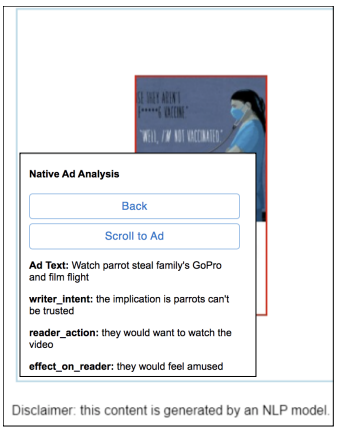

Figure 1: Web browser application with Misinfo Reaction Frames backend model for implication prediction.

The second aim of my research is ensuring factual groundedness of text generation. Hallucinatory behavior of generative models poses an ongoing challenge to using machine-generated implications for misinformation detection, fact-checking and online interventions [9, 10]. Model hallucinations arise from parameterization of models learned during training being reliant only on training data distributions, rather than real-world plausibility or verified factual knowledge. This can be mitigated by methods like loss truncation [11] or adding controllability and flexibility to language models through architectures like my work on generator-discriminator frameworks [12]. However, I found that **the metrics by which we evaluate quality of generative models are themselves a bottleneck to improving factuality**. I developed a theoretically-grounded meta-evaluation, Go Figure, for evaluating the metrics used to assess performance of text summarization models. I found a number of limitations in both common summarization metrics and recently proposed metrics that use end-task objectives like

[1] https://www.media.mit.edu/people/cynthiab/overview/
[2] https://www.media.mit.edu/

question-answering to rank factuality [9]. I also defined a taxonomy of errors based on the distribution found in summaries generated by the T5 language model [6], and uncovered that performance varies across factual error types. This points out (1) the risk of machine-generated disinformation being encouraged in selection of generative models due to incorrectly specified evaluation, and (2) the need for more fine-grained schema in assessing factuality to avoid overweighting easy-to-detect errors.

Mitigation of Harms from Neural Language Models:

The third aim of my broader research agenda is ensuring algorithms are designed in a way that mitigates social inequities. In 2019, collaborators and I released a paper revealing that existing Twitter hate speech detection datasets contained significant bias towards speakers of African-American Vernacular English (AAVE) [13]. This bias in training data is amplified by text classifiers like those underlying the Google Perspective API (See Figure 2 for label distributions),[3] leading to a concerning situation where at-risk populations are being discriminated against by deployment of the very algorithms developed for their protection. This finding highlighted two major fairness issues within the current framing of toxic language detection (and more broadly, NLP tasks) - **disembodiment** and **imbalanced evaluation**.

As noted by [14], NLP models and datasets become disembodied when researchers attempt to abstract away context in pursuit of universality. Curation of large-scale datasets means stripping away important features like information about online ecosystems in which conversations arise or audience vs. speaker dynamics. Such pragmatic context is critical for tasks like toxic language detection (e.g. when the identity of the speaker may change the perceived intent and harmfulness of their speech). The harms of interpreting data out-of-context may be mitigated by designing evaluations based on real-world data distributions, where the concerns of stakeholders are properly considered as part of the model selection process. Our work highlighted that both in research and production models were not thoroughly evaluated with respect to real-world use and the diverse perspectives of their users.

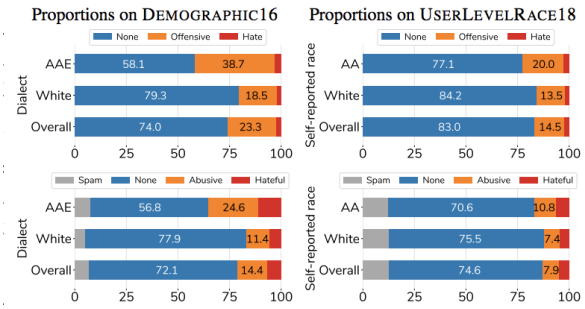

Figure 2: Average probability mass of toxicity classes as given by hate speech classifiers trained on Twitter data. Proportions are shown for AAVE, White-aligned English, and overall.

To address these issues, collaborators and I induced a large language model to produce the types of data missing from current datasets. This led to the Toxigen project, in which **we view large pretrained language models like GPT-3 [15] as tools to extract web data and distill knowledge about socially biased and harmful online text** [16]. For Toxigen, we use carefully crafted prompts to encourage GPT-3 to generate either "harmless" or "toxic" language directed at one of 13 at-risk groups commonly targeted by hate speech. Toxigen also uses an adversarial variant of beam search decoding in which GPT-3 is encouraged to generate text that fools a hate speech

[3] https://perspectiveapi.com/

classifier by starting with a prompt with a specific label (e.g. toxic language) and maximizing a combination of the GPT-3 language model likelihood and the classifier likelihood of a contradictory label (e.g. harmless language).

In three years, these works have already been influential. The Risk of Racial Bias paper was covered by multiple media outlets including Forbes[4] and was nominated for a best short paper at ACL 2019. A follow-up work that I co-authored, Social Bias Frames [17], won best paper at the WeCNLP 2020 summit. Toxigen was covered by TechCrunch[5] in 2022 and will be used in production for content filtering at Microsoft Research.

Future Directions

For my long-term research plan to build equitable and factually correct systems, I intend to merge these frameworks for interpreting potential harm of language with ongoing work exploring information disorder and factuality. I found in my work on Toxigen that a significant percentage of machine-generated language made a factual claim, and postulate that there may be benefit from taking a holistic view of online harm in which we consider the role social biases play in disinformation spread and vice versa. I am currently working with researchers at UW, AI2, and CMU to develop a unified multi-modal framework for detection of false or harmful language.

Based on my prior work, I have identified two areas I plan to explore to further my main research goals:

Improving explainability of language models: My work has shown that the ability of models to reason correctly can be improved by making problem-solving operations explicit [18]. This has the added benefit of making decision-making more explainable and trustworthy to users. Explanation generation can also allow human users to understand implications conveyed by text. As a long-term goal, I plan to use machine-generated explanations for refining users' ability to read with discernment and question the veracity and underlying intent of communications.

In-the-wild testing and strategies for distributional robustness: A pervasive issue in NLP research is model brittleness to out-of-distribution data [19], which is particularly problematic for large-scale and in-the-wild deployment of algorithms. I have work recently accepted to Findings of EMNLP on generating naturalistic adversaries for improving robustness of text classifiers. In my future work, I plan to focus on development of best practices for simulating in-the-wild testing and evaluations that consider a diverse pool of potential stakeholders using online crowdsourcing. This builds upon previous user-centric work I have done, to design effective crowdsourcing interfaces for data collection (e.g. for Toxigen) and user-focused model validation (e.g. A/B testing of machine-generated Misinfo Reaction Frames implications). I would also like to work with journalists and grassroots organizations (e.g. DAIR Institute,[6] Masakhane,[7] the Anti-Defamation League[8]) to ensure groundedness and generalization of research.

Publications

https://www.forbes.com/sites/nicolemartin1/2019/08/13/googles-artificial-intelligence-hate-speech-detector-is-racially-biased/?sh=6bccb821326c

https://techcrunch.com/2022/05/23/microsoft-claims-its-new-projects-make-language-models-safer-to-use/

https://www.dair-institute.org/

https://www.masakhane.io/

https://www.adl.org/

[1] Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 127 (November 2019), 26 pages. https://doi.org/10.1145/3359229

[2] Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo Reaction Frames: Reasoning about Readers' Reactions to News Headlines. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3108–3127, Dublin, Ireland. Association for Computational Linguistics.

[3] Charles J. Fillmore. 1976. Frame Semantics and the Nature of Language. Annals of the New York Academy of Sciences, 280.

[4] Ian Apperly. 2011. Mindreaders: The cognitive basis of "theory of mind." Psychology Press.

[5] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nick Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. AAAI.

[6] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21, 1, Article 140 (January 2020), 67 pages.

[7] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. Unpublished manuscript.

[8] Emily Dinan, Gavin Abercrombie, A. Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022. SafetyKit: First Aid for Measuring Safety in Open-domain Conversational Systems. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4113–4133, Dublin, Ireland. Association for Computational Linguistics.

[9] Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A Meta Evaluation of Factuality in Summarization. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 478–487, Online. Association for Computational Linguistics.

[10] Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing Faithfulness in Knowledge-Grounded Dialogue with Controllable Features. In 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 704–718, Online. Association for Computational Linguistics.

[11] Daniel Kang and Tatsunori B. Hashimoto. 2020. Improved Natural Language Generation via Loss Truncation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 718–731, Online. Association for Computational Linguistics.

[12] Saadia Gabriel, Antoine Bosselut, Jeff Da, Ari Holtzman, Jan Buys, Kyle Lo, Asli Celikyilmaz, and Yejin Choi. 2021. Discourse Understanding and Factual Consistency in Abstractive Summarization. In Proceedings of the 16th Conference of the European Chapter of the Association for

Computational Linguistics: Main Volume, pages 435–447, Online. Association for Computational Linguistics.

[13] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

[14] Zeerak Waseem, Joachim Bingel and Isabelle Augenstein. 2021. Disembodied Machine Learning: On the Illusion of Objectivity in NLP. ArXiv abs/2101.11974.

[15] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei. 2020. Language Models are Few-Shot Learners. In Proceedings of NeurIPS.

[16] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

[17] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5477–5490, Online. Association for Computational Linguistics.

[18] Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

[19] Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. Measure and Improve Robustness in NLP Models: A Survey. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4569–4586, Seattle, United States. Association for Computational Linguistics.

[20] Saadia Gabriel, Chandra Bhagavatula, Vered Shwartz, Ronan Le Bras, Maxwell Forbes and Yejin Choi. 2021. Paragraph-Level Commonsense Transformers with Recurrent Memory. Proceedings of AAAI.

[21] Commonsense Dialogic Question Generation. 2022. Pedro Colon-Hernandez, Saadia Gabriel, Yejin Choi, Cynthia Breazeal and Hae Won Park. In Submission.

[22] Saadia Gabriel, Hamid Palangi, Yejin Choi. 2022. NaturalAdversaries: Can Naturalistic Adversaries Be as Effective as Artificial Adversaries. Findings of EMNLP.