Akshita Jha
https://akshitajha.github.io
Email : akshitajha@vt.edu
LinkedIn: linkedin.com/in/akshitajha/

Education

- **Virginia Tech** — Washington DC-Baltimore Area
  Ph.D. Candidate, Dept. of Computer Science; GPA: 4.00 — *Aug. 2019 – Present*
  - **Thesis**: Enhancing Robustness and Mitigating Stereotypes at Scale for Generative Models
  - **Committee**: Chandan K. Reddy, Lifu Huang, Xuan Wang, Su Lin Blodgett, Vinodkumar Prabhakaran
- **IIIT-Hyderabad** — Hyderabad, India
  MS by Research, Natural Language Processing
- **IIIT-Hyderabad** — Hyderabad, India
  Bachelor of Technology (Hons.), Computer Science

Research and Work Experience

- **Google Research** — Mountain View, CA
  Research Intern, Responsible AI — *May 2023 - August 2023*
  - **Safety of Text-to-Image Models**: Conducted a global scale analysis of visual stereotypes in Text-to-Image generations by building a multi-faceted evaluation paradigm for detecting regional stereotypes in generated images of 135 identity groups around the globe. Used for evaluating image generations from models like Stable Diffusion.
- **Google Research** — Mountain View, CA
  Research Intern, Responsible AI — *August 2022 - January 2023*
  - **Safety of Large Language Models (Foundation Models)**: Built a broad-coverage stereotype benchmark dataset by leveraging generative capabilities of PaLM, and GPT-3 using few-shot prompting. Evaluated several large language models (LLMs) for the presence of regional stereotypes and leveraged a globally diverse rater pool to validate their prevalence in society – both in-region and out-region.
- **AI Lab, InterDigital** — Palo Alto, CA
  Research Intern — *May 2020 - August 2020*
  - **Interpretable Long Document Matching**: Built a contrastive learning framework to compute (dis)similarity within and across different chunks and sections of long documents in an interpretable manner using contrastive learning along with BERT embeddings and custom position embeddings.
- **Virginia Tech** — Arlington, VA
  Research Assistant, Advisor: Chandan K. Reddy — *August 2019 - Present*
  - **Distinguishing between task-specific flaws and (un)fairness of generative models**: Large-scale evaluation of foundation models to identify whether the observed stereotypical disparities arise from inherent biases against identity groups (unfairness) or from the model's intrinsic (in)ability to (in)effectively handle a downstream task.
  - **Robustness of Code Programming Language Models**: Designed CodeAttack, a black-box attack model that uses code structure to generate effective, efficient, and imperceptible adversarial code samples. Successfully demonstrates the vulnerabilities and evaluates the robustness of programming language models at scale to code-specific adversarial attacks on code translation, code repair, and code summarization.
  - **Challenges in Transformer-based Models for Long-Form Document Matching**: Demonstrate the effectiveness of simple neural models and simple embeddings over transformer-based models on the task of document matching. We show that simple models are at par with the more complex BERT-based models while taking significantly less training time, energy, and memory.
  - **Fair Representation Learning**: Built a model for learning fair disentangled representations while ensuring the utility of the learned representation for downstream tasks.
- **University of Michigan** — Ann Arbor, MI
  Research Assistant, Advisors: David Jurgens, Libby Hemphill — *June 2018 - September 2018*
  - **Measuring Benevolent Prejudice from Twitter Data** : Automatically identify cross-cultural biases against different social and political groups in Twitter data using clustering algorithms.
  - **De-escalation of Hate-Speech on Reddit**: Built a bot to automatically detect the escalation of hostility in a Reddit thread and automatically post a 'de-escalating' comment to prevent the conversation from going off-rails.

Selected Publications

For the most recent list, please refer to my Google Scholar.

- **Akshita Jha**, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Chandan K. Reddy, Sunipa Dev. *"ViSAGE: A Global-Scale Analysis of Visual Stereotypes in Text-to-Image Generation"*, Under Review
- **Akshita Jha**, Aida Davani, Shachi Dave, Vinodkumar Prabhakaran, Sunipa Dev. *"SeeGULL: A Stereotype Benchmark with Broad Geo-Cultural Coverage Leveraging Generative Models"*, ACL 2023
- Sunipa Dev, **Akshita Jha**, Jaya Goyal, Dinesh Tewari, Shachi Dave, Vinodkumar Prabhakaran. *"Building Stereotype Repositories with Complementary Approaches for Scale and Depth"*, Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), EACL 2023
- **Akshita Jha** and Chandan K. Reddy. *"CodeAttack: Code-based Adversarial Attacks for Pre-Trained Programming Language Models"*. AAAI 2023. **[Spotlight Presentation]**
- **Akshita Jha**, Adithya Samavedhi, Vineeth Mohan, and Chandan K. Reddy. *"Transformer based Models for Long Document Comparison: Challenges and Empirical Analysis"*, The 17th Conference of EACL (Findings), 2023
- **Akshita Jha**, Vineeth Mohan, Jaideep Chandrashekar, Adithya Samavedhi, and Chandan K. Reddy. *"Supervised Contrastive Learning for Interpretable Long-Form Document Matching"*. ACM Transactions on KDD, May, 2022.
- **Akshita Jha**, Bhanukiran Vinazamuri & Chandan K. Reddy. *"Fair Representation Learning using Interpolation Enabled Disentanglement"*. 2021
- **Akshita Jha** and Radhika Mamidi. *"When does a Compliment become Sexist? Analysis and Classification of Ambivalent Sexism using Twitter Data"*. Proceedings of the Second Workshop on Natural Language Processing and Computational Social Science, ACL 2017. **[Spotlight Presentation]**

Awards

- **Travel Awards**: AAAI 2023, ACL 2023, EACL 2023
- **CS Research Mentorship Program Scholar**: Google 2021
- **Grace Hopper Scholarship**: Was a Grace Hopper Scholar for the year 2020
- **Member of Phi Kappa Phi Honor Society**: Awarded to top 1% for academic excellence.

Service

- Reviewer for several machine learning (ML), artificial intelligence (AI), and Natural Language Processing (NLP) conferences and workshops like ACL, AAAI, EACL, NAACL, KDD, AIES, etc., and journals like TKDD, and IEEE.

Programming Skills

- **Languages and Technologies**: Python, PyTorch, TensorFlow, Keras

Relevant Coursework

- Machine Learning, Deep Learning, Advanced Machine Learning, Topics in Human-Computer Interaction, Natural Language Processing, Data Mining.