AKSHITA JHA

akshitajha@vt.edu | https://akshitajha.github.io/

EDUCATION

Virginia Tech, Washington DC-Baltimore Area *Fall 2019 - Present*
Ph.D. Candidate, Dept. of Computer Science GPA: 4.0/4.0
Advisor: Dr. Chandan K. Reddy

IIIT-Hyderabad, Hyderabad, India
MS by Research, Computational Linguistics
Advisor: Dr. Radhika Mamidi

IIIT-Hyderabad, Hyderabad, India
Bachelor of Technology (Hons.), Computer Science GPA: 8.11/10.00

RESEARCH INTERESTS

Large Language Models, Fairness and Bias, Robustness, Natural Language Processing, Generative AI

WORK EXPERIENCE

Responsible AI, Google Research, CA | *Research Intern* *May 2023 - Aug 2023*
Responsible AI, Google Research, CA | *Research Intern* *Aug 2022 - Jan 2023*
AI Lab, InterDigital, Palo Alto, CA | *Research Intern* *May 2020 - Aug. 2020*
Google Summer of Code, Debian | *Intern* *May 2015 - Aug. 2015*
The Linux Foundation, OpenDaylight | *Deb Maintainer* *May 2016 - Apr 2017*
Intuit, Bangalore, India | *Software Engineer 1* *Aug 2017 - May 2018*

PUBLICATIONS

- **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
- [Spotlight Presentation] **Akshita Jha** and Chandan K. Reddy. *"CodeAttack: Code-based Adversarial Attacks for Pre-Trained Programming Language Models"*. AAAI Conference on Artificial Intelligence, AAAI 2023.
- **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
- [Spotlight Presentation] **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.

RESEARCH PROJECTS

Disentangled Representations of Fairness
Collaborators: Chandan K. Reddy, Virginia Tech; Bhanukiran Vinzamuri, IBM Research
- Proposed a deep learning model that learnt latent representations to effectively disentangle factors of variation that help in identifying features for protected attributes. The model used interpolation along with adversarial autoencoder for better disentanglement.

Detecting Ambivalent Sexism from Twitter Data
Collaborators: Radhika Mamidi, IIIT-Hyderabad
- Created a novel dataset of 10,000 sexist tweets (Benevolent Sexist, Hostile Sexist and Others). Built a classifier using Support Vector Machines (SVM), Sequence-to-Sequence models, and FastText to classify tweets into one of the above three classes.

Measuring Benevolent Prejudice from Twitter Data
Collaborators: David Jurgens, University of Michigan, Ann Arbor
- Automatically generated words on a semantic axis in a high dimensional vector space using k-nearest neighbors, breadth-first search and dependency parsing. The aim was to identify implicit and explicit cross-cultural biases against different social and political groups in Twitter data by using inherent degrees of warmth and competence.

De-escalation of Hate-Speech on Reddit
Collaborators: Libby Hemphill, David Jurgens, University of Michigan, Ann Arbor
- Built a bot to automatically detect the escalation of hostility in a Reddit thread using a vocabulary of insults, part-of-speech tagger and dependency parser and automatically post a 'de-escalating' comment to prevent the conversation from going off the rails.

AWARDS AND HONORS

- **Travel Award** for AAAI, 2023; EACL, 2023; ACL, 2023
- **CS Research Mentorship Program Scholar**, Google, 2021
- **Grace Hopper Celebration Scholarship**, 2020
- **Member of the Phi Kappa Phi Honor Society for Academic Excellence**

SKILLS

Languages (Proficient in): Python
Languages (Familiar with): C, C++, Java
Tools/Libraries: PyTorch, Keras, LaTeX, Bash Shell Scripting, Git

RELEVANT COURSEWORK

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