

Danish Pruthi

Applied Scientist
Amazon AI

danish@hey.com
+1 (412) 819-8046

Research Interests

I research in the areas of natural language processing and deep learning. Specifically, my research addresses core issues concerning the interpretability of deep learning models by devising methods and evaluation protocols that improve and quantify stakeholders' understanding of such models. Going forward, I am interested in developing mechanisms for users to better control and update models; and applying my research to applications that could have a large societal impact.

Education

- 2016–2021 PhD in Language & Information Technologies
Carnegie Mellon University
Advisors: Graham Neubig & Zachary C. Lipton
Thesis Committee: William W. Cohen & Michael Collins (+ advisors)
- 2016–2018 Masters in Language Technologies
Carnegie Mellon University
- 2011–2015 B.E. (Hons.) in Computer Science
Birla Institute of Technology and Science, Pilani (BITS Pilani)

Professional Experience

- 2020 Fall Student Researcher, Google Research, Pittsburgh (Host: William W. Cohen)
- 2020 Summer Research Intern, Google Research, Pittsburgh (Host: William W. Cohen)
- 2019 Summer Research Intern, Facebook AI Research (FAIR), New York (Host: Brenden Lake)
- 2015–2016 Project Assistant, Indian Institute of Science (IISc), Bangalore (Host: Partha Talukdar)
- 2015 Spring Research Intern, Microsoft Research, Bangalore
- 2014 Summer Software Engineering Intern, Google, Hyderabad

Selected Awards

- 2018–2019 CMU Presidential Fellowship
- 2017–2018 Siebel Scholarship (USD 35,000)
- 2019 Best Demo Runner-up Award at NAACL 2019
- 2011 KVPY Scholarship (declined to pursue engineering)

Publications

Total citations: 655, h-index: 10, i-index: 10 (Google Scholar: <http://bit.ly/danish037>)

Refereed Journal Papers

- [1] Evaluating Explanations: How much do explanations from the teacher aid students?
Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins,
 Zachary C. Lipton, Graham Neubig, William W. Cohen
 Transactions of the Association for Computational Linguistics (TACL, 2021)

Refereed Conference Papers

- [1] Explain, Edit, and Understand: Rethinking User Study Design for Evaluating Model Explanations
 Siddhant Arora*, Danish Pruthi*, Norman Sadeh, William W. Cohen, Zachary C. Lipton, Graham Neubig
 Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI, 2022)
- [2] Do Context-Aware Translation Models Pay the Right Attention?
 Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F. T. Martins, Graham Neubig.
 The Annual Meeting of the Association for Computational Linguistics (ACL, 2021)
- [3] Weakly- and Semi-supervised Evidence Extraction
Danish Pruthi, Bhuwan Dhingra, Graham Neubig, Zachary C. Lipton
 Findings of Empirical Methods in Natural Language Processing (Findings of EMNLP, 2020)
- [4] Why and when should you pool? Analyzing Pooling in Recurrent Architectures
 Pratyush Maini, Keshav Kolluru, Danish Pruthi, Mausam
 Findings of Empirical Methods in Natural Language Processing (Findings of EMNLP, 2020)
- [5] Learning to Deceive with Attention-Based Explanations
Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, Zachary C. Lipton
 The Annual Meeting of the Association for Computational Linguistics (ACL, 2020)
- [6] Combating Adversarial Misspellings with Robust Word Recognition
Danish Pruthi, Bhuwan Dhingra, Zachary C. Lipton
 The Annual Meeting of the Association for Computational Linguistics (ACL, 2019)
- [7] Simple and Effective Semi-Supervised Question Answering
 Bhuwan Dhingra*, Danish Pruthi*, Dheeraj Rajagopal*
 Meeting of the North American Chapter of the ACL (NAACL, 2018)

- [8] SPINE: SParse Interpretable Neural Embeddings
Danish Pruthi^{*}, Harsh Jhamtani^{*}, Anant Subramanian^{*}, Taylor Berg-Kirkpatrick, Eduard Hovy
AAAI Conference on Artificial Intelligence (AAAI, 2018)
- [9] Discovering Response Eliciting Factors in Social Question Answering: A Reddit Inspired Study
Danish Pruthi, Yogesh Dahiya, Partha Talukdar
AAAI Conference on Web and Social Media (ICWSM, 2016)
- [10] Maxxyt: An Autonomous Wearable Device for Real-time Tracking of a Wide Range of Exercises
Danish Pruthi, Ayush Jain, KrishnaMurthy Jatavallabhula, Ruppesh Nalwaya, and Puneet Teja
International Conference on Modelling and Simulation. (UKSim, 2015)

^{*} denotes equal contribution

Refereed System Demonstrations

- [1] NeuSpell: A Neural Spelling Correction Toolkit
Sai Muralidhar Jayanthi, Danish Pruthi, Graham Neubig
Conference on Empirical Methods in Natural Language Processing (EMNLP, 2020)
- [2] compare-mt: A Tool for Holistic Comparison of Language Generation Systems
Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, Xinyi Wang, John Wieting
Meeting of the North American Chapter of the ACL (NAACL, 2019)
Recipient of the Best Demo Runner-up Award

Under Review

- [1] The Impact of Vocabulary Size on Robustness of NLP Models
Pratyush Maini, Zico Kolter, Zachary C. Lipton, Danish Pruthi
Submitted to the Conference on Empirical Methods in Natural Language Processing (EMNLP, 2022)
- [2] Learning to Scaffold: Optimizing Model Explanations for Teaching
Patrick Fernandes, Marcos Treviso, Danish Pruthi, André F. T. Martins, Graham Neubig
Submitted to the Conference on Neural Information Processing Systems (NeurIPS, 2022)
- [3] Measures of Information Reflect Memorization Patterns
Rachit Bansal, Danish Pruthi, Yonatan Belinkov
Submitted to the Conference on Neural Information Processing Systems (NeurIPS, 2022)

Grants

Contributed to preparation of the following grants:

- [1] Expert-in-the-Loop Neural Summarization for Consequential Domains
National Science Foundation (NSF) Medium Research Award
(Amount: USD 583,746; PI: Zachary Lipton; Timeline: 2022–2026)
- [2] Robustifying NLP by Exploiting Invariances Learned via Human Interaction
Facebook Research Award
(Amount: USD 80,000; PI: Zachary Lipton; Timeline: 2019–2020)

Research Mentoring

Mentored 10 students, spanning 6 past and 2 ongoing projects. Past projects have resulted in 5 papers so far.

- Siddhant Arora (MS at CMU)
- Rachit Bansal (B.Tech at DTU)
- Pratyush Maini (B.Tech from IIT D; now PhD at CMU)
- Patrick Fernandes (PhD at CMU)
- Sai Muralidhar Jayanthi (MS at CMU)
- Kayo Yin (MS at CMU)
- Dev Seth (BE at Duke University)
- Punit S. Koura (MS at CMU)
- Vivek Pandit (MS at CMU)
- Divyansh Agarwal (MS at CMU)

Teaching Experience

- Teaching Assistant for Neural Networks for NLP (11-747)
My responsibilities included preparing a lecture on interpretability, creating quizzes, and mentoring student projects.
- Teaching Assistant for Introduction to Machine Learning (PhD) (10-701)
My responsibilities included conducting recitations, holding office hours, creating and grading assignments.
- Competitive Programming Special Interest Group (CPSIG)
Led the special interest group at BITS Pilani. Delivered lectures spanning data structures, algorithms, graph theory and game theory. Conducted similar workshops in sister campuses of BITS Goa and BITS Hyderabad.

Invited Talks & Panels

- Evaluating Model Explanations
Conference on Deployable AI, March 2022
Allen Institute for AI, January 2022
Amazon, December 2021
Google AI, November 2021
- Towards Model Understanding
IISc Bangalore, April 2022
IIT Bombay, October 2021
IIT Delhi, October 2021
IIT Madras, October 2021
Unbabel AI Seminar, September 2021

- **A Tale of Evidence and Explanations**
Data Science Seminar at University of Utah, December 2020
Machine Learning Seminar at Twitter Inc., October 2020
NLP Weekly at Google AI, July 2020
- **Attention and its Interpretation**
IIT Delhi, January 2020
ACL, July 2020
- **Model Interpretation**
Alumni Research Talk at BITS Pilani, January 2020
Guest Lecture for Neural Networks for NLP course at CMU, Spring 2019, 2020, 2021
Guest Lecture for Computational Semantics Course at CMU, Spring 2019
- **Interpreting Word Representations**
Indian Institute of Science Bangalore, January 2018
Student Research Symposium at CMU, August 2017 (**Honorable mention for the best presentation**)
- **Panel Discussion on Higher Education for Undergraduates**
Session for undergraduate student researchers at ACL, July 2020
IIT Jodhpur, November 2020
- **Panel Discussion on Life at LTI & CMU**
Open house for admitted students at CMU, Spring 2020, 2019

Professional Services

- **Area Chair**
2022: EMNLP
2021: ICLR Workshop on Responsible AI
- **Reviewing**
2021: ACL, NAACL, FAccT, JAIR
2020: EMNLP (**Outstanding Reviewer**), ACL, ICLR, AAAI
2019: EMNLP, NAACL
- **Volunteering**
2021: Faculty Hiring Committee
2020: Graduate Application Support Program

Selected Press

- [1] It's Too Easy to Hide Bias in Deep-Learning Systems.
IEEE Spectrum
- [2] How to Measure The Performance Of Explainability Models.
Analytics India Magazine
- [3] 12 Interesting Papers From ACL 2020.
Analytics India Magazine

References

- [1] **Zachary C. Lipton (advisor)**
BP Junior Chair Assistant Professor
Carnegie Mellon University
Email: zlipton@cs.cmu.edu

- [2] **Graham Neubig (advisor)**
Associate Professor
Carnegie Mellon University
Email: gneubig@cs.cmu.edu

- [3] **William W. Cohen**
Professor, Carnegie Mellon University
Principal Scientist, Google Research
Email: wcohen@google.com

- [4] **Partha Talukdar**
Associate Professor (on leave), IISc Bangalore
Staff Research Scientist, Google Research
Email: ppt@iisc.ac.in

- [5] **Michael Collins**
Vikram S. Pandit Professor of Computer Science
Columbia University
Senior Staff Scientist, Google Research
Email: mjcollins@google.com