

SIDDHANT ARORA

Phone : +1-724-680-6603

Email : siddhana@andrew.cmu.edu

Github : <https://github.com/siddhu001>

RESEARCH INTERESTS

Interpretability, Spoken Language Understanding, Robustness, Machine Learning on Structured Text

EDUCATION

Carnegie Mellon University, Pittsburgh

Aug. 2020 - Aug. 2022

Master of Science, Language Technology (Funded with Fellowship)

CQPA : 4.21/4.00

School of Computer Science.

Indian Institute of Technology, Delhi

Jul. 2015 - Jul. 2020

B.Tech and M. Tech (Dual Degree)

CGPA: 9.428/10 , Department Rank: 1

Department of Computer Science and Engineering.

PUBLICATIONS

1. **ESPnet-SLU: Advancing Spoken Language Understanding through ESPnet:** [Siddhant Arora](#), Siddharth Dalmia, Pavel Denisov, Xuankai Chang, Yushi Ueda, Yifan Peng, Yuekai Zhang, Sujay Kumar, Karthik Ganesan, Brian Yan, Ngoc Thang Vu, Alan W Black, Shinji Watanabe. Under submission at ICASSP 2022 Conference
2. **Explain, Edit, and Understand: Rethinking User Study Design for Evaluating Model Explanations:** [Siddhant Arora](#), Norman Sadeh, William W. Cohen, Zachary C. Lipton, Graham Neubig, Danish Pruthi. Under submission at AAAI 2022 Conference
3. **Rethinking End-to-End Evaluation of Decomposable Tasks: A Case Study on Spoken Language Understanding:** [Siddhant Arora](#)*, Alissa Ostapenko*, Vijay Viswanathan*, Siddharth Dalmia*, Florian Metze, Shinji Watanabe, Alan Black. Interspeech 2021 Conference [URL]
4. **IterefinE: Iterative KG Refinement Embeddings using Symbolic Knowledge:** [Siddhant Arora](#), Srikanth Bedathur, Maya Ramnath, Deepak Sharma. AKBC 2020 Conference [URL]
5. **Capreolus: A Toolkit for End-to-End Neural Ad Hoc Retrieval:** Andrew Yates, [Siddhant Arora](#), Xinyu Zhang, Wei Yang, Kevin Martin Jose, Jimmy Lin. WSDM 2020 Demo Track [URL]
6. **A survey on Graph Neural Network for Knowledge Graph Completion:** [Siddhant Arora](#). Arxiv 2020. [URL]
7. **Investigating Ad-Hoc Retrieval Method Selection with Features Inspired by IR Axioms :** [Siddhant Arora](#), Andrew Yates. AMIR Workshop, ECIR 2019, Cologne, Germany [URL]
8. **On Embeddings in Relational Database:** [Siddhant Arora](#) and Srikanth Bedathur. KR2ML Workshop, NeurIPS 2019 [URL]
9. **Understanding Community Rivalry on Social Media: A Case Study of Two Footballing Giants :** Sopan Khosla, [Siddhant Arora](#), Abhilash Nandy, Ankita Saxena, Anandhavelu N. HUMANIZE Workshop, Intelligent User Interfaces (IUI) Conference 2019. [URL]
10. **GATCheck: A Detailed Analysis of Graph Attention Networks:** Lovish Madaan*, [Siddhant Arora](#)*. Mining and Learning with Graph Workshop, KDD 2020 (Accepted)

ACADEMIC ACHIEVEMENTS

- **Top 10% of MLT class** in CMU in first 2 semesters
- **CMU Graduate Research Fellowship**
- **Institute Silver Medal** for securing Department Rank 1 in Dual Degree Program of Computer Science and Engineering at IIT Delhi
- **Merit certificate** at IIT Delhi in 9 out of 10 semesters given to only top 7% students
- All India Rank **11 (amongst 1,200,000 candidates) in JEE Mains 2015** and **289 (amongst 140,000 candidates) in IIT-JEE Advanced 2015**
- **KVPY Scholar fellowship** (300 selected amongst 600,000 candidates) in January 2014

RESEARCH EXPERIENCE

Graduate Research Assistant

Prof. Shinji Watanabe

CMU Pittsburgh

May 2021 - Current

Open Source Benchmark for Spoken Language Understanding

- Devised the toolkit ESPnet-SLU, which is designed for quick development of spoken language understanding models inside ESPnet.
- Provided implementations that cover all experiment processes for intent classification, slot filling, emotion recognition and dialogue acts classification datasets using various ASR and NLU submodules.
- Released an open-source toolkit and the trained models that match or even significantly outperform the SOTA by upto 8% on SLU benchmarks.

Graduate Research Assistant

Prof. Graham Neubig, Prof. Zachary Lipton and Prof. William Cohen

CMU Pittsburgh

Feb. 2021 - Aug. 2021

User Study Design for Evaluating Explanations

- Published Human Intelligence Tasks on Amazon MTurk, where users interact with deception detection models that have been trained to distinguish between genuine and fake hotel reviews.
- Proposed a design where users have query access to the model and can observe how model predictions and explanations change in real time.
- Investigated extending the simulation task by prompting users to edit examples to reduce the model confidence towards the predicted class.

Graduate Research Assistant

Prof. Norman Sadeh

CMU Pittsburgh

Aug. 2020 - Current

Multilingual Annotation for Privacy Documents

- Devised an annotation scheme designed to capture the nuances of new privacy regulations, namely GDPR in EU and CCPA/CPRA in US.
- Collected the first bilingual corpus of privacy policies in English and German.
- Developed classifiers to automatically extract privacy disclosures and find differences between English and German policy.

Graduate Research Assistant

Prof. Shinji Watanabe and Prof. Alan Black

CMU Pittsburgh

Nov 2020 - Mar 2021

Robust Test Sets for Spoken Language Understanding

- Proposed a framework to construct robust test sets over sub-task specific utility functions.
- Observed more actionable comparisons between different architectures.
- Released our splits and tools for the community: <https://MASEeval.github.io/>.

Undergraduate Research Assistant*Prof. Srikanta Bedathur*

IIT Delhi

*Jul 2019-Jun 2020***Learning Embeddings on Relation Databases**

- Used embedding to enable semantic queries on relational database.
- Leveraged semantic level information to embed continuous data, utilized LSTM for time series, experimented with attention models (Transformers) and KG embedding models.

Undergraduate Research Assistant*Prof. Srikanta Bedathur*

IIT Delhi

*Sep 2019-Mar 2020***Knowledge Graph Refinement**

- Utilized probabilistic soft logic for removing noise from Knowledge graph based on ontology and Knowledge Graph embedding for inferring new facts together in a feedback manner.
- Observed that our framework produced upto 9% improvement in Knowledge Graph completion and refinement over a range of benchmarks.

Undergraduate Research Intern*Prof. Andrew Yates*

MPII Saarbrücken

*Jan 2019 - May 2019***Stress Testing IR Models**

- Combined IR models based on instance level query features that are inspired by IR axiom.
- Designed toolkit with implementations of prominent neural ranking integrated with the Anserini toolkit.
- Performed stress testing of neural IR models based on IR axioms.

Undergraduate Research Intern*Prof. Tom Gedeon*

ANU Canberra

*Nov 2017 - Dec 2017***Emotion Recognition**

- Used pretrained CNN model as feature extractor to differentiate between posed & actual smiles.
- Investigated using user gaze data along with image features in Extreme Learning Machine(ELM).

WORK EXPERIENCE**Research Intern***Adobe Systems*

Bangalore, KA, India

*May 2018 - Jul 2018***Understanding Toxic Content Online**

- Performed hate speech detection & analysed how events impact hate content with synthetic time series.
- Characterised behaviour of key instigators of online hate speech.
- Predicted potential hate instigators for events to prevent toxic content from spreading online.

Data Scientist*Bidgely Inc.*

Mountain View, CA, US

*May 2019 - Jul 2019***Energy Disaggregation for Household**

- Created electric vehicle disaggregation by applying CNN on energy heatmaps for households.
- Performed solar disaggregation to estimate solar generation from net energy load and topic modelling on user reviews to extract customer feedback and grievances.

TEACHING EXPERIENCE**Teaching Assistant***Prof. Shravan Kumar*

IIT Delhi

*Jan 2019- May 2019***MTL100 course (Calculus)**

- Conducted weekly tutorial sessions to discuss class material and help with solving assignment problems.

- Developed weekly assessments and performed grading of quizzes and exam questions.

Teaching Assistant

Prof. Subodh Kumar

COL106 (Data Structures)

IIT Delhi

Jul 2019-Dec 2019

- Conducted weekly lab sessions to discuss programming assignments and clarify student questions.
- Designed automated tests to grade programming assignments.

Teaching Assistant

Prof. Rijurekha Sen

COL380 (Parallel and Distributed Programming)

IIT Delhi

Jul 2019-Dec 2019

- Created programming assignments and developed automated tests for their assessment.
- Conducted vivas to test knowledge on programming assignments and course material.

INVITED TALKS & PANELS

- “Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text”. **FTC Privacy Con.** June 2021.
- “Using NLP to Understand Differences in Privacy Disclosures and Protections under Different Jurisdictions”. **Cylabs Lab Tour.** October 2021.

VOLUNTEER WORK

Volunteer

Prof. Rajesh Prasad

Aarohan Programme

IIT Delhi

Oct 2015

- Guided underprivileged children on exam preparation for engineering entrance exams.
- Taught Physics and clarified their questions about the exam pattern.

Volunteer

Prof. Rajesh Prasad

National Association of Blind

IIT Delhi

May-Jun 2016

- Volunteered in the National Association of Blind for 12 hrs/week.
- Helped to organize Annual Diwali Fair and also guided differently-abled students in the undergraduate admissions process.

Volunteer

Prof. Sangeeta Kohli

Academic Mentor

IIT Delhi

May-Jun 2016

- Volunteered as Academic mentor for the first-year undergraduate students at IIT Delhi.
- Taught basic syntax about programming in C++ and held doubt clarifying sessions in student hostels.

HOBBIES AND INTEREST

- Traveling- Fortunate to travel to more than 10 countries. Trip to Tokyo, Japan, was the most memorable one. Participated in the Dai Ginza Matsuri festival and tried the most delicious yakisoba and mochi dessert.
- Cricket- Have been passionate about cricket since childhood.
- Swimming- Love to go for a swim on the weekends.