VIBHOR AGARWAL

+44-7880334078 ◇ agarwalvibhor84@gmail.com ◇ vibhor98.github.io ◇ www.linkedin.com/in/vibhor98

EDUCATION

University of Surrey July 2021 - Nov 2024
Doctor of Philosophy in Natural Language Processing *Surrey, England, UK*

The LNM Institute Of Information Technology August 2016 - June 2020
Bachelor of Technology in Computer Science and Engineering *Jaipur, India*

Overall GPA: 9.45/10

WORK EXPERIENCE

Nokia Bell Labs Dec 2024 - Present
Research Scientist *Cambridge, UK*

· I am working in the Responsible AI team at the intersection of Natural Language Processing and AI safety to minimize harms and biases in LLMs and make them transparent and trustworthy.

Roku Jul 2024 - Sep 2024
Applied Science Internship *Cambridge, UK*

· I worked on developing an LLM-powered task-oriented Chatbot for Smart Home users.

JP Morgan AI Research Sep 2023 - Dec 2023
NLP Research Internship *London, UK*

· I worked on detecting and combating hallucinations in code generated by Large Language Models.

Georgia Institute of Technology Jul 2023 - Aug 2023
Visiting Researcher *Atlanta, Georgia, USA*

· I worked on automatically detecting and combating hallucinations in Large Language Models for healthcare.

Queen Mary University of London Oct 2022 - Apr 2023
Visiting Researcher *London, UK*

· I worked on understanding online conversations in decentralized platforms such as Mastodon and Pleroma and studied how these conversations evolve and federate over different instances. I also proposed various moderation strategies for hate speech in decentralized conversations.

Rewire Online Jul 2022 - Sep 2022
NLP Research Internship *London, UK*

· I worked on a research project, funded by DSO Singapore, to detect online abuse in code-switched languages such as Singlish and low-resource languages such as Malay and Indonesian.
· Crawled multilingual dataset of online abuse from Reddit and developed machine learning models to detect abusive language.

The Alan Turing Institute Jul 2021 - Sep 2021
Visiting Researcher, Data Study Group *London, UK*

· In this Data Study Group, organized by Turing Institute and University of Leeds, UK, I worked on a research project to predict Vet-AI's demands, which is an AI powered research company for Pets' healthcare.

Media.net (Directi) Jan 2020 - Jun 2021
Site Reliability Engineer *Mumbai, India*

- Media.Net is a leading contextual advertising tech company.
- As full-time employee, I led lots of projects. The largest among them was to optimise the Machine Learning pipeline of a Keyword Recommender model and integrate it with TensorFlow Extended.
- The Keyword Recommender model, using the context of a webpage, suggests keywords for which ads can be shown on webpages.

SLK Software May 2019 - Jul 2019
Software Engineering Internship *Bangalore, India*

- Built a cross-platform desktop application using Electron and ReactJS to analyse Java applications.
- Given the project path as input, the application can generate its corresponding Class Diagram and Data-Flow Diagram for class analysis and properties like inheritance. It also finds the dead code, unused variables, duplicate imports and other problems in the project.
- Technologies used - JavaScript (ReactJS, Electron, Node.js), Python, and MongoDB.

Google Summer of Code 2018 — The Oppia Foundation May 2018 - Aug 2018
Open Source Developer *Remote*

- Oppia aims at building interactive platform for both lesson creators and learners and brings learning freely available in the interesting and story telling way.
- Implemented two new interactions – "Number with units" and "Drag and drop sort" to enhance the learning experience of the students and help creators in making effective lessons.
- Technologies used - JavaScript (AngularJS, JQuery), Python, and Google Cloud.

RESEARCH PUBLICATIONS

GraphNLI: A Graph-based Natural Language Inference Model for Polarity Prediction in Online Debates ACM Web Conference (WWW) 2022
Vibhor Agarwal*, Sagar Joglekar, Anthony P. Young, Nishanth Sastry*

- Developed a novel Graph-based deep learning model to predict the argumentative relations of *attack* and *support* (polarities) in online debates.
- GraphNLI leverages neighboring context in online discussions through graph walks, along with the local context, and has wide applications in NLP field, such as Recognizing Textual Entailment.

Conversation Kernels: Retrieval-Augmented Context Modeling for Online Conversation Understanding ICWSM 2025
Vibhor Agarwal*, Arjoo Gupta, Suparna De, Nishanth Sastry*

MedHalu: Hallucinations in Healthcare Queries by Large Language Models ICWSM 2026
Vibhor Agarwal*, Yiqiao Jin, Mohit Chandra, Munmun De Choudhary, Nishanth Sastry, Srijan Kumar*

AnnoBERT: Effective Representation of Multiple Annotators' Perspectives and Label Semantics for Hate Speech Detection ICWSM 2023
Wenjie Yin,* ***Vibhor*** *Agarwal*, Aiqi Jiang*, Arkaitz Zubiaga, Nishanth Sastry*

- Model different annotators' perspectives using Collaborative Topic Regression in subjective tasks such as Hate Speech detection.
- AnnoBERT fuses annotator perspectives and label semantics to improve Hate Speech detection.

A Graph-Based Context-Aware Model to Understand Online Conversations ACM Transactions on the Web (TWEB) 2023 Journal
Vibhor Agarwal*, Anthony P. Young, Sagar Joglekar, Nishanth Sastry*

GASCOM: Graph-based Attentive Semantic Context Modeling for Online Conversation Understanding Online Social Networks and Media (OSNEM) 2024 Journal
Vibhor Agarwal*, Yu Chen, Nishanth Sastry*

- We propose GASCOM and design two novel *semantic-aware graph-based conversation context selection algorithms* for retrieving relevant context nodes which consider both the graph structure and semantic meanings of the conversation context.
- We further design a *token-level multi-head graph attention mechanism* to pay different attentions to different tokens from the selected context utterances for fine-grained conversation context modeling.

HateRephrase: Zero- and Few-Shot Reduction of Hate Intensity in Online Posts using Large Language Models Under Review
Vibhor Agarwal*, Yu Chen, Nishanth Sastry*

Decentralised Moderation for Interoperable Social Networks: A Conversation-based Approach for Pleroma and the Fediverse ICWSM 2024
Vibhor Agarwal*, Aravindh Raman, Nishanth Sastry, Ahmed Sayed, Gareth Tyson, Ignacio Castro*

CodeMirage: Hallucinations in Code Generated by Large Language Models AutoMates@IJCAI 2024
Vibhor Agarwal*, Yulong Pei, Salwa Alamir, Xiaomo Liu*

Improving the Detection of Multilingual Online Attacks with Rich Social Media Data from Singapore ACL 2023
Janosch Haber, Bertie Vidgen, Matthew Chapman, ***Vibhor Agarwal****, Roy Ka-Wei Lee, Yong Keong Yap, Paul Röttger*

AI in the Gray: Exploring Moderation Policies in Dialogic Large Language Models vs. Human Answers in Controversial Topics CIKM 2023
Vahid Ghafouri, ***Vibhor Agarwal****, Y. Zhang, N. Sastry, J. Such, G. Suarez-Tangil*

Biases and Ethical Considerations for Machine Learning Pipelines in the Computational Social Sciences Ethics in Artificial Intelligence: Bias, Fairness and Beyond (Book Chapter)
Suparna De, Shalini Jangra, ***Vibhor Agarwal****, Jon Johnson, Nishanth Sastry*

"Way back then": A Data-driven View of 25+ years of Web Evolution WWW 2022
Vibhor Agarwal*, Nishanth Sastry*

- An exploratory data analysis and measurement study to quantify how the Web has evolved, by leveraging historical data crawled from *Archive.org*, *Google Trends*, and *Alexa*.

Under the Spotlight: Web Tracking in Indian Partisan News Websites ICWSM 2021
Vibhor Agarwal*, Yash Vekaria, Pushkal Agarwal, Sangeeta Mahapatra, Shounak Set, Sakthi Balan Muthiah, Nishanth Sastry, Nicolas Kourtellis*

Differential Tracking Across Topical Webpages of Indian News Media WebSci 2021
Yash Vekaria, ***Vibhor Agarwal****, Pushkal Agarwal, Sangeeta Mahapatra, Sakthi Balan, Nishanth Sastry, Nicholas Kourtellis*

AWARDS AND ACHIEVEMENTS

- Program Committee member at WWW, AAAI ICWSM and Reviewer at WSDM, CIKM, and COLING.
- Delivered research talks at British Telecom (Responsible AI), Georgia Tech, N. Carolina State University, King's College London, and Queen Mary University of London.
- Presented my poster at **ELLIS Cambridge Summer School** in July 2022.
- Received **Ultimate Web Science Quiz Award** at ACM WebSci 2021.
- Received funded **PhD studentship** at University of Surrey, UK.

- Received the **Best Bachelor's Thesis Award** in the graduated class of 2020 at LNMIIT, Jaipur, India.
- **Quarter finalist in India Innovation Challenge Design Contest' 2018** organised by Department of Science and Technology and IIM Bangalore for our robust 'Pro-active Video Surveillance System'.
- **Awarded Meritorious Certification and Scholarship** for securing position in top 1 percentile at LNMIIT, Jaipur.
- Awarded **Certificate of Appreciation** for **Google Summer of Code' 18** at LNMIIT, Jaipur.
- Earned **Silver Medal** at **State level Mental ability competition (Abacus)**.

TECHNICAL SKILLS

Programming Languages	Python, JavaScript, C, Java, MATLAB
Artificial Intelligence	Machine Learning, NLP, Large Language Models, Deep Learning
Tools	PyTorch, Keras, Scikit-Learn, Numpy, Pandas, Matplotlib, and Seaborn
Frameworks	Flask (Python), AngularJS, ReactJS, Electron, and Node.js (JavaScript)
Database	MySQL, SQLite, MongoDB and Redis (NoSQL)
Version Control	Git and GitHub

OPEN SOURCE

Data Version Control (DVC) Jul 2019 - Nov 2019

- DVC is an open-source version control for Data Science and Machine Learning projects.
- Mostly contributed to the technical documentation of the project and made bug fixes in the code.

The Oppia Foundation Sep 2017 - Jul 2019

- Contributed to this great Google's open source project for almost two years.
- Was part of core developer's team, implemented new features and helped new Geeks in getting started with the basics. Was also a **maintainer** keeping track of the Pull Requests.

POSITIONS OF RESPONSIBILITY

Teaching Assistant, University of Surrey, UK Sep 2021 - June 2024

- Served as a Teaching Assistant in NLP, Computational Intelligence, Advanced Web Technologies, and Databases labs and helped Master's students in understanding and implementing relevant CS concepts.

Data Science Lead — Computer Society of India (LNMIIT Chapter) Aug 2018 - May 2019
Computer Society of India (CSI) is the first and the largest body of computer professionals in India.

- Guided engineering students in accomplishing Data Science projects and conducted various workshops and research talks related to state-of-the-art algorithms in ML and Data Science.

Google Code-In 2018 Mentor May 2018 - Aug 2018

- Selected as the mentor in GCI 2018 program to guide school students to contribute in open source projects under **JBoss (RedHat)** organization.