Mohna Chakraborty

Department of Computer Science, Iowa State University
716-748-5386 | mohnac@iastate.edu | https://mohna0310.github.io

RESEARCH INTERESTS

The primary objective of my post-doctoral research is to investigate the practical applications of Generative AI (GenAI) techniques, with a particular focus on enhancing the capabilities of large language models (LLMs) to effectively navigate social interactions. My work aims to develop innovative methods that improve the explainability, interpretability, replicability, and overall robustness of LLMs in real-world social contexts.

During my doctoral studies, I concentrated on advancing the fields of review analysis, information extraction using weak supervision, data mining, natural language processing, and machine learning. My research addressed the critical challenge of the scarcity of labeled textual data by devising methods that streamline the annotation process, minimizing human effort while ensuring that these approaches are practical for everyday systems without the need for costly hardware. This focus on accessibility and efficiency has been a driving force in my work.

Throughout my academic journey, I have made significant contributions to the field, with six published works in prestigious conferences such as ACL, UAI, SIGKDD, and ESEC/FSE, as well as in reputable workshops like RANLP and the ML Reproducibility Challenge. Additionally, I have several papers currently under review, reflecting my ongoing commitment to advancing the state of the art in these domains.

EDUCATION

University of Michigan — Ann Arbor, MI, USA
Post Doctorate (MIDAS Data Science Fellow) — *Sep 2024 - Sep 2026*
- **Research Area:** Generative AI, LLMs, Data mining, Machine Learning, Natural Language Processing

Iowa State University — Ames, IA, USA
Ph.D in Computer Science — *Aug 2020 - Aug 2024*
- **Research Area:** Data mining, Machine Learning, Natural Language Processing

State University of New York at Buffalo — Buffalo, NY, USA
M.S in Computer Science — *Aug 2018 - July 2020*
- **Research Area:** Data Mining, Machine Learning, Natural Language Processing

PUBLICATIONS

Conferences

ACL, UAI, SIGKDD, NeurIPS, TKDD, RANLP, ESEC/FSE, ML Reproducibility Challenge

- **Mohna Chakraborty***, Adithya Kulkarni*, Qi Li, "Zero-shot Approach to Overcome Perturbation Sensitivity of Prompts.", **61st Annual Meeting of the Association for Computational Linguistics (ACL), 2023**, url: https://aclanthology.org/2023.acl-long.313. (Acceptance Rate: 20.8%)
- Adithya Kulkarni, **Mohna Chakraborty**, Sihong Xie, Qi Li, "Optimal Budget Allocation for Crowdsourcing Labels for Graphs.", **39th Conference on Uncertainty in Artificial Intelligence (UAI), 2023**, url: https://proceedings.mlr.press/v216/kulkarni23a.html. (Acceptance Rate: 29.3%)
- **Mohna Chakraborty***, Adithya Kulkarni*, Qi Li, "Open-Domain Aspect-Opinion Co-Mining with Double-Layer Span Extraction.", **28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD), 2022**, url: https://doi.org/10.1145/3534678.3539386. (Acceptance Rate: 14.9%)
- Abhishek Kumar Mishra*, **Mohna Chakraborty***, "Does local pruning offer task-specific models to learn effectively?.", **Student Research Workshop Associated with 13th International Conference on Recent Advances in Natural Language Processing RANLP, 2021**, url: https://aclanthology.org/2021.ranlp-srw.17.
- **Mohna Chakraborty**, "Does reusing pre-trained NLP model propagate bugs?.", **29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE), 2021**, url: https://dl.acm.org/doi/10.1145/3468264.3473494.

- Richard D Jiles, **Mohna Chakraborty**, "[Re] Domain Generalization using Causal Matching.", **ML Reproducibility Challenge, 2021**, url: https://openreview.net/forum?id=r43elaGmhCY.
- Adithya Kulkarni, **Mohna Chakraborty**, Qi Li. "Budget Allocation Exploiting Label Correlation between Instances.", (Under review) Neural Information Processing System: NeurIPS 2024.
- **Mohna Chakraborty**, Adithya Kulkarni, Qi Li. "Weakly Supervised Open-Domain Aspect-based Sentiment Analysis.", (Under review) Transactions on Knowledge Discovery in Data: TKDD 2023.

WORK EXPERIENCE

Data Science Intern May 2023 - July 2023
Home Depot *Atlanta, GA, USA*

- I worked on building a **sentence-based transformer personalized search ranker** using personalized signals to drive better engagements.
- The current search ranker in production uses a **tree-based model** which does not catch semantic similarity between the search query and product and does not consider personalized information about the customers. Therefore, the need was to create a generalizable modeling framework that can handle multi-modality and understand the correlation between different customers to mitigate the cold start problem.
- The proposed model improved MAP@8 by 20.38% MRR@8 by 21.27% NDCG@8 by 16.31% compared with the PROD baseline. Thus, improved relevancy for top-ranked products (i.e., higher precision).

Data Science Intern May 2022 - Aug 2022
Epsilon Data Management, LLC *Chicago, IL, USA*

- The current production system uses **Spark SQL**, which does not support atomic operations like **upsert, delete**, etc., so the system overwrites the entire table for every update, resulting in higher resource consumption and time.
- **Apache Hudi** is proposed as an alternative tool to solve issues like upsert and delete operations. We tested multiple workflows using Apache HUDI, including parameter tuning to validate its effectiveness by testing real-life scenarios with 1 billion data for up to 10% upsert operations, and the result shows a 37% increase in run time compared with Spark SQL.

Data Science Intern May 2021 - Aug 2021
Epsilon Data Management, LLC *Chicago, IL, USA*

- The ability to accurately classify individual names plays a crucial role in the quality of the final product. Yet this ability is hampered due to heterogeneity in data collection and validation.
- Current production methods validate the name data using **rule-based approaches**, limiting its ability to update or scale. Therefore, to alleviate this problem, we propose using **machine learning algorithms** on top of rule-based features and encoded features with 191 million data.

- Based on the classifiers' performance, **Random Forest** achieved a 91% F1 score with **oversampling** and **customized features**, explaining the need to incorporate better features to help the learning model better and faster.

Data Analyst Intern May 2019 - March 2020
Delaware North *Buffalo, NY, USA*

- Worked on training a **Logistic Regression model** to predict passenger occupancy across the US Airport and used its prediction to train another Logistic Regression model that predicts the number of transaction counts and labor force needed during various days in various kiosks or restaurants across airport terminals.

* Worked with **Beautiful Soup** to web scrape data like attendance, duration of the game, home and opponent team details, and other details for the games like NBA, NHL, NFL, and MLB from their official website, and used the data to train a **predictive model** to understand the trend of the crowd for all these games.

Junior Data Scientist Sep 2015 - Dec 2017
Ericsson India Global Pvt. Ltd *Mumbai, MH, India*

* Two years of work experience as a Junior Data Scientist in an **agile environment** with hands-on experience in designing and implementing Machine Learning Algorithms.
* Utilized alarms to perform **Anomaly detection** that captures unusual site behavior of base stations via **supervised and unsupervised machine learning models**. The developed models significantly reduced the alarms by 30%. The models helped prevent the faults from happening through early predictions and proactive decisions.
* Developed a **regression model** to identify the cause of weak network connection by performing a descriptive analysis to gain insights into the dataset, summary statistics, and analyzing features impacting the target correlation among variables. The developed regression model achieved a 10% more precise prediction than the previous year.

HONORS AND AWARDS

- **MINK WIC Conference Poster Competition:** Best Research Poster Award for my paper "Zero-shot Approach to Overcome Perturbation Sensitivity of Prompts" (accepted at ACL, 2023) at MINK WIC Conference. (October, 2023)
- **MINK WIC (Missouri, Iowa, Nebraska, Kansas Women in Computing):** Selected to represent Iowa State University at MINK WIC (An ACM celebration of Women in Computing). (October, 2023)
- **Guest Lecturer Invitation:** I have taught a graduate-level course at Iowa State University as a guest lecturer for COM S 571X (Responsible AI: Risk Management in Data Driven Discovery.) to teach the students about representation learning, transformer models, and future research directions to develop trustworthy machine learning methods for natural language processing.
- **Research Presentation Competition:** Second place at the 7th Annual Research Day Competition in the Computer Science Department at Iowa State University. (May, 2023)
- **Grace Hopper Celebration:** Selected to represent Iowa State University for the prestigious and competitive Grace Hopper Celebration. (September, 2022)
- **Travel Award:** One among 46 students selected by SIGKDD for the student travel award among all the applicants. (August, 2022)
- **Research Presentation Competition:** First place at the 6th Annual Research Day Competition in the Computer Science Department at Iowa State University. (May, 2022)

REFERENCES

Name	Title	Email	Department	University
Qi Li	Assistant Professor	qli@iastate.edu	Computer Science	Iowa State University
Wallapak Tavanapong	Professor	tavanapo@iastate.edu	Computer Science	Iowa State University