Mohna Chakraborty

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

RESEARCH INTERESTS

My research interests are in the domain of data mining, natural language processing, and machine learning. My work focuses on solving the problem of the scarcity of labeled textual data. My proposed approaches can facilitate the annotation process with minimal human effort and can be implemented in daily-use systems without expensive hardware, promoting accessibility to everyone. Through my research, I have contributed several key methods in top conferences and workshops. I have **six published works** in top conferences like **ACL, UAI, SIGKDD, ESEC/FSE** and workshops like **RANLP and ML Reproducibility Challenge**. Below, I summarize my research interests:

Data annotation using pre-trained large language models

Extracting useful information from unstructured and abundant reviews is pivotal for the success of businesses. Through my projects, I have proposed several unsupervised (zero-shot prompt-based) and weakly supervised methods to label the important information in the reviews. The weak labels obtained using my proposed approaches can have noise and bias. Therefore, I have proposed an approach to tackle the issues of noise and bias in the weak labels so that machine learning models can be properly trained using the generated weak labels. The research projects resulted in **two publications** in highly competitive NLP and data mining conferences, such as **ACL, 2023 and SIGKDD, 2022**.

Large sizes of LLMs restrict their applicability to devices with high computational resources since they cannot be deployed on edge devices with limited computational resources. To overcome this challenge, I investigated different methods of unstructured pruning on task-specific models to compress these large language models. The project resulted in **one publication** in **RANLP, 2021** workshop.

Data annotation using crowd workers

Data annotation is costly. Samples in the unlabeled corpora may have label correlations, and utilizing this correlation can help reduce data annotation costs. Obtaining crowd labels for a few correlated samples can help infer labels for the remaining correlated samples. Through my research, I have been working on projects to take advantage of this correlation between samples in graph structures to choose important samples to obtain crowd labels and then infer labels for the remaining correlated samples by propagation labeling information. The research projects have resulted in **one publication** in **UAI, 2023**, a highly competitive artificial intelligence conference.

EDUCATION

Iowa State University — Ames, IA, USA
Ph.D in Computer Science — *Aug 2020 - Aug 2024 (expected)*

- **Research Area:** Data mining, Machine Learning, Natural Language Processing
- **Advisor:** Dr. Qi Li

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
- **Advisor:** Dr. David Doermann
- **Thesis Title:** Using Machine Learning for Predicting Aspect-Wise Satisfaction Ratings by Semantic Analysis of Text

West Bengal University of Technology, Kolkata — Kolkata, WB, India
B.Tech in Electronics and Communication Engineering — *June 2011 - May 2015*

PUBLICATIONS

Conferences

ACL, UAI, SIGKDD, RANLP, ESEC/FSE, ML Reproducibility Challenge, EMNLP

- **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.

HONORS AND AWARDS

- **MINK WIC (Missouri, Iowa, Nebraska, Kansas Women in Computing) 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)

CONFERENCE PRESENTATIONS

- Open-Domain Aspect-Opinion Co-Mining with Double-Layer Span Extraction, SIGKDD, 2022, Washington DC.
- Does reusing pre-trained NLP model propagate bugs? ESEC/FSE, 2021 (virtual conference presentation)

SERVICE

- Served as a program committee member for HCOMP' 2022, and EMNLP' 2022 and as a review member for PAKDD' 2022, and EACL'2023 conferences.
- I serve as a mentor for an undergraduate student at Iowa State University for the McNair program. The program encourages low-income, first-generation, and underrepresented undergraduates to pursue graduate studies.

TEACHING

Teaching Assistant
Iowa State University *Ames, IA, USA*

COM S 309 - Software Engineering Practices *Fall 2020, Spring 2021, Fall 2021 and Spring 2022*

- ∗ This is a core, mandatory undergraduate course that introduces methods for managing software development and developing a large software from scratch to the end product with Android client, Spring back-end, concurrent features, etc.
- ∗ I have graded, and closely supervised 10 project teams (40+ students) from a class of 300+ student.
- ∗ In three out of four semesters, my supervised teams won the best project award in the whole class.

COM S 363 - Introduction to Database Management Systems *Spring 2022 and Fall 2023*

- ∗ This course is an undergraduate course covering basic topics of database systems.
- ∗ I created, mentored, and graded the Assignments/Quizzes/Exams of the students in this course and others.

COM S 227 - Object Oreinted Programming *Spring 2023*

- This course is an undergraduate course covering basic topics of Object Oriented Programming. I worked as a head teaching assistant for this course.
- I created, mentored, and graded the Assignments/Quizzes/Exams of the students in this course and others.

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.