Aisha Khatun

+1 226 899 8911 | aisha.khatun@uwaterloo.ca | linkedin.com/in/tanny411
github.com/tanny411 | Google Scholar | tanny411.github.io | Kitchener, ON, Canada

PROFESSIONAL SUMMARY

- Data Scientist and Researcher with 4+ years of experience in Predictive Analytics, Machine Learning, and Natural Language Processing.
- Adept at handling large datasets and applying cutting-edge technology, like Generative AI, to derive impactful data-driven business insights.
- Demonstrated ability to effectively collaborate within diverse teams as well as work independently to drive projects from start to finish, showcasing adaptability.
- Ability to proficiently convey complex concepts to non-technical audiences through presentations and storytelling.

SKILLS

Languages: Python, SQL, SPARQL, C++ (Proficient); Javascript, Java, Scala, PHP (Comfortable)

Tools: Keras, PyTorch, Fastai, Tensorflow 2, HuggingFace, Numpy, Pandas, Spark, Hadoop, Airflow, Tableau, Power BI

Other Tech: HTML/CSS, JQuery, NodeJS, ReactJS, Flask, MongoDB, PostgreSQL, Git, GitLab, BitBucket, Gerrit, Phabricator

EXPERIENCE

Research Data Scientist Feb 2023 - Present
Wikimedia Foundation, USA (Remote)

- Addressed deployment bottlenecks in Wikipedia link recommendation system by creating a language-agnostic model that replaced 300+ individual language-dependent models.
- Increased the standard of Wikipedia articles through automated copy-editing by extracting and analyzing Wiktionary data and detecting commonly misspelled words in 100+ language Wikipedias.
- Used Python, PySpark, Hadoop, Sklearn, and XGBoost.

Graduate Research Student Sep 2022 - Present
University of Waterloo, Canada

- Analyzed the capabilities and limitations of a wide range of LLMs in responding to sensitive statements (e.g., stereotypes, conspiracy theories, etc), consistency across settings, and robustness to prompt variations.
- Created a dashboard to help select appropriate LLMs for specific business use cases based on fact-checking, bias detection, and instruction-following abilities across 37 open and closed-source models.
- Used HuggingFace, Tableau, and LLM APIs.

Data Analyst Apr 2021 - Aug 2022
Wikimedia Foundation, USA (Remote)

- Informed business decisions to reduce data import speed and query timeouts in Wikidata Query Service by analyzing Wikidata to find large and most frequently queried subgraphs.
- Enabled graph and query statistics monitoring by creating a pipeline that calculates and saves metrics periodically.
- Used Spark, SQL, and Airflow.

Outreachy Intern (Data Science) Dec 2020 - Mar 2021
Wikimedia Foundation, USA (Remote)

- Chosen as one of 54 Outreachy Interns from 1000+ applicants for outstanding Open Source contributions.
- Reduced redundancy in Wikipedia Lua modules by co-creating a tool that identifies unique modules for centralization in Abstract Wikipedia through data analysis and source code similarity metrics.

Machine Learning Engineer Mar 2020 - Sep 2020
Therap BD Ltd, Bangladesh
- Improved in-office attendance application with high-accuracy face detection in image and video footage.
- Optimized care home efficiency by implementing an OCR system that extracts measurements from pulse oximeter images for swift COVID-19 detection.
- Increased care home safety and accessibility by developing an ML-based fall detection system using inertial sensor readings from smartwatches.

Research Assistant Nov 2018 - Jun 2020
SUST NLP Lab, Bangladesh
- Advanced Bengali NLP's foundation by curating several large datasets and training versatile language models with various tokenization methods.
- Improved Authorship Attribution by curating the largest dataset in Bengali Literature, developing deep learning architectures, and analyzing the effects of tokenization and pre-training datasets on downstream tasks.

EDUCATION

Masters of Mathematics in Computer Science (Thesis) Aug 2022 - April 2024
University of Waterloo, Waterloo, Canada

Bachelor of Science in Computer Science and Engineering Jan 2016 - Mar 2020
Shahjalal University of Science and Technology (SUST), Sylhet, Bangladesh

PUBLICATIONS

- **Khatun, A.**, & Brown, D. *A Study on Large Language Models' Limitations in Multiple-Choice Question Answering.* ArXiv, 2024.
- **Khatun, A.**, & Brown, D. *Reliability Check: An Analysis of GPT-3's Response to Sensitive Topics and Prompt Wording.* TrustNLP Workshop, ACL 2023.
- **Khatun, A**. et al. *Authorship Attribution in Bangla Literature (AABL) via Transfer Learning using ULMFiT***.** ACM Journal, ACM TALLIP 2022.
- **Khatun, A**. et al. *A Subword Level Language Model for Bangla Language*. Springer Conference, IJCCI 2020.
- **Khatun, A**. et al. *Authorship Attribution in Bangla literature using Character-level CNN*. IEEE Conference, ICCIT 2019.

PROJECTS

- **Image Captioning** - Merged CV and NLP techniques to develop an image captioning deep neural network from scratch. Used Pytorch to leverage pre-trained models for text and image inputs.
- **Online Study Portal** - Developed a social networking web app to make group interaction easier and organized using the LAMP stack. Features include authentication, group file storage and organization, instant messaging, file searching, notifications, posting, and commenting.
- **Others** - Facial keypoint detection, Flight delay detection, Food review, and FMNIST classification.

VOLUNTEER WORK

Mentor and Project Lead Jan 2024 - April 2024
Directed Reading Program, University of Waterloo
- Mentored undergraduate students in a generative AI project, guiding them through the process of learning about LLMs, conducting literature reviews, and applying LLMs for computational creativity.

Workshop Instructor Jan 2023
Women in Computer Science, University of Waterloo
- As an instructor, led a hands-on AI workshop at the Women in Computer Science Conference (WiCSCon), guiding participants through an introduction to data analysis, ML algorithms, and best practices.

Coordinator Feb 2022 - August 2022

Google Summer of Code and Outreachy, Wikimedia Foundation

- Co-organized and led events throughout the internship period to welcome interns, connect them to existing Wikimedians, and solve issues as they came along.
- Ensured intern engagement and retention by documenting progress and hosting AMAs, among other things.

AWARDS

- Barbara Hayes-Roth Award for Women in Math and Computer Science, 2023.
- Received awards in several Programming Contests, notably the National Girls Programming Contest (2017 and 2018) and ACM ICPC ASIA Regional Programming Contest, Dhaka Site 2018.
- National Scholarship awarded by the Education Board, Government of Bangladesh, 2020.