Mohammad Sabik Irbaz

GitHub: github.com/msi1427 **Email: mirbaz@gmu.edu**
Portfolio: msi1427.github.io/ **Phone:** +17034593856

I am a Ph.D. student at George Mason University, specializing in **natural language processing with a focus on medical applications**. I am currently working as a **Graduate Research Assistant** under the supervision of **Dr. Kevin Lybarger**. I earned my BSc in Computer Science and Engineering from the Islamic University of Technology, Bangladesh, in 2021. After graduation, I spent **2.5 years** in the data science industry, serving as a Data Scientist at Leadbook and as Lead Machine Learning Engineer at Omdena, where I led end-to-end data science projects. My research areas include but are not limited to **text simplification, text style transfer, and clinical machine learning**.

Research Interests

I am primarily excited about Natural Language Processing (NLP), Artificial Intelligence (AI), Computational Social Science, and Data Science research. My research interests include:
- AI for Good (Social, Clinical & Other Disciplines) **[P1, P2, P5, P6, P7, P8, P9, P10]**
- NLP Applications (Low-resource NLP, QA, Commonsense Reasoning, etc.) **[P3, P4]**
- Fact verification using historical data **[P7]**
- Data-centric AI **[P6]**

Education

George Mason University — Fairfax, Virginia, USA
Ph.D. in Information Technology (INFT) — Aug, 2023 - May, 2028 (*expected*)
CGPA: 4.00/4.00
Supervisor: Dr. Kevin Lybarger

Islamic University of Technology(IUT) — Dhaka, Bangladesh
B.Sc. in Computer Science and Engineering (CSE) — Jan, 2017 - Mar, 2021
CGPA: 3.44/4.00
Dissertation: “Medical Expertise Style Transfer using Denoising Autoencoders” [Thesis Book] [Presentation]
Supervisor: Dr. Md. Kamrul Hasan and Dr. Hasan Mahmud

Research Experience

{$^{\alpha}$primary contributor; $^{\beta}$secondary contributor; $^{\gamma}$supervisor}

Lybarger Lab — Aug, 2023 - Present

Graduate Research Assistant
Department of IST, George Mason University, Virginia, USA
Advisor(s): Dr. Kevin Lybarger
(1) $^{\alpha}$Health Text Simplification and Novel Strategies for RL — *NLP, Data Scraping, LLMs, RLHF* **[P9]**
(2) $^{\alpha}$OCR and Vision-LLMs for IE from Handwritten Violence Reports — *OCRs, Multimodal LLMs* **[P10]**

Fordham Human Centered AI Research Lab — Mar, 2021 - June, 2023

Department of CIS, Fordham University, New York, USA
Advisor(s): Dr. Md. Ruhul Amin
(1) $^{\alpha}$COVID-19 Public Policy Conformity Analysis — *NLP, Data Extraction, Topic Modeling, NLI, EDA* **[P7]**
(2) $^{\beta}$Mitigating Social Bias in NLP — *NLP, Explainability, Text Style Transfer* **[P8]**

Systems and Software Lab (SSL) — Jul, 2019 - June, 2023

Department of CSE, IUT, Dhaka, Bangladesh
Advisor(s): Dr. Md. Kamrul Hasan, Dr. Hasan Mahmud
(1) $^{\alpha}$Medical Expertise Text Style Transfer — *NLP, Active Learning, Text Style Transfer* **[P6]**

Computer Vision Lab (CVLab) Nov, 2020 - June, 2023

Department of CSE, IUT, Dhaka, Bangladesh
Advisor(s): Sabbir Ahmed
(1) $^{\alpha}$Automatic Radiology Report Generation — *NLP, Image Captioning, Image Segmentation* [Report]
(2) $^{\gamma}$Logical Reasoning QA using Standardized Test Questions — *NLP, QA, CSR, Data Collection, NLI*

Reviewing Experience

Reviewed multiple conference papers for **EMNLP 2022** and **WACV 2023**

Publications

More updated list here or Google Scholar
{*equal contribution, ^{c}corresponding author}

[P10] **Mohammad Sabik Irbaz**, Md. Mushfiqur Rahman, Katherine Scafide, David Lattanzi, Janusz Wojtusiak, Kevin Lybarger^c. **"Optical Character Recognition for Information Extraction from Handwritten Assault and Violence Reports"**. *Poster Presentation at AMIA 2024 Symposium Workshop: Using Data Science and AI to Advance Violence and Injury Prevention.* [*Manuscript in Preparation for journal submission.*]

[P9] Md Mushfiqur Rahman*, **Mohammad Sabik Irbaz***, Kai North, Michelle S Williams, Marcos Zampieri, Kevin Lybarger^c. **"Health Text Simplification: An Annotated Corpus for Digestive Cancer Education and Novel Strategies for Reinforcement Learning"**. *Journal of Biomedical Informatics, Volume 158, October 2024*

[P8] Md. Shihab Shahriar, **Mohammad Sabik Irbaz**, and Md. Ruhul Amin^c. **"Mitigating Social Bias in NLP using explainable AI methods"**. *Manuscript in Preparation.*

[P7] **Mohammad Sabik Irbaz**, Rupak Sarkar, Md. Ruhul Amin^c and Ashiqur R. KhudaBukhsh. **"COVID-19 Public Policy Conformity Analysis"**. *Manuscript in Preparation.* [Presentation]

[P6] **Mohammad Sabik Irbaz**, Abir Azad, Zarif Ahmed, Md. Ashfaq Raiyan, Raian Rahman, Hasan Mahmud, and Md. Kamrul Hasan^c. **"c-MSD: A Curated Dataset for Medical Expertise Style Transfer using Active Learning Approach"**. *Under Review.*

[P5] **Mohammad Sabik Irbaz**^{c}, Fardin Ahsan Sakib, and Lutfun Nahar Lota. **"Predicting User-specific Future Activities using LSTM-based Multi-label Classification"**. *Proceedings of the 4th International Conference on Activity and Behavior Computing 2022* [Presentation]

[P4] Alif Ashrafee*, Akib Mohammed Khan*, **Mohammad Sabik Irbaz**^{c}, and MD Abdullah Al Nasim. **"Real-time Bangla License Plate Recognition Pipeline for Low Resource Video-based Applications"**. *Real-World Surveillance: Applications and Challenges Workshop, collocated with IEEE/CVF WACV 2022* [Presentation] [Poster]

[P3] Fahim Shahriar Khan, Mueeze Al Mushabbir, **Mohammad Sabik Irbaz**^{c}, and MD Abdullah Al Nasim. **"End-to-End Natural Language Understanding Pipeline for Bangla Conversational Agents"**. *Proceedings of the 20th IEEE International Conference on Machine Learning and Applications, 2021* [Presentation]

[P2] **Mohammad Sabik Irbaz**^{c}, MD Abdullah Al Nasim, and Refat E Ferdous. **"Real-Time Face Recognition System for Remote Employee Tracking"**. *Proceedings of the International Conference on Big Data, IoT, and Machine Learning, 2022. Springer, Singapore.*

[P1] **Mohammad Sabik Irbaz***, Abir Azad*, Tanjila Alam Sathi, and Lutfun Nahar Lota^c. **"Nurse Care Activity Recognition Based on Machine Learning Techniques Using Accelerometer Data"**. *Proceedings of the 8th International Workshop on Human Activity Sensing Corpus and Applications, collocated with ACM UbiComp/ISWC 2020* [DOI] [Presentation]

Industry Experience

Leadbook Pte. Ltd. — Singapore

Data Scientist — Dec, 2021 - Feb, 2023

Supervising and maintaining the following projects to improve our Lead Generation process
(1) ***Industry Type Classification:*** According to LinkedIn, there are 148 types of industries and one company can be attributed to multiple industries. We hired 10 annotators and trained them to annotate 30K data to facilitate multi-label classification. Our final model is trained in 3 steps. a) Language Model (LM) finetuning with 1M company data, b) Knowledge distillation from large zero-shot classifier c) Using the finetuned LM and distilled classifier to finetune on the 30K annotated data.
(2) ***Job Title Clustering:*** We have millions of contacts in our database. An important problem is to cluster contact job titles into categories in an unsupervised or semi-supervised manner. Most of my works involves improving and maintaining this pipeline.The project involves — NLP, Data Cleaning, Clustering, Phrase Similarity [Code]
(3) Some of my regular responsibilities involve maintaining data quality and catering to **customized lead generation** requests utilizing different data transformation, scraping and ML methods.

Omdena Inc. — New York, USA

Lead Machine Learning Engineer — Apr, 2021 - Jun, 2021

Collaborated with 50 ML and data engineers from 20 countries on **Equilo Project** - Building Gender Equality Assessment Functionality using NLP. This large scale work was divided into 7 major tasks. I was the **task leader** of the team building the **recommender system** for the organizations with low gender equality scores. I also lead the team building the web app and deploying the whole data and ML pipeline to **GCP**.
This project involved — *Scraping, NLP, GCP, Firestore, Django, ReactJS, StreamLit, Cloud SQL* [Article]

Pioneer Alpha Ltd. — Dhaka, Bangladesh

Lead Machine Learning Engineer — Jan, 2021 - Oct, 2021

Supervised and delivered the following core projects successfully to the corresponding clients:
(1) Multilingual Chatbot as Business Assistant — *NLP, NLU, Low-resource, Tensorflow, Rasa, Flask* [Code] **[P3]**
(2) Low-resource Bangla License Plate Detection and Recognition — *OCR, Tensorflow, OCR, Flask* [Code] **[P4]**
(3) Face Detection and Recognition for Remote Office — *Vision, OpenCV, dlib, Flask* **[P2]**
(4) Trello Automation for Directed Premiere Video Editing — *Vision, OpenCV, dlib, Adobe Premiere, Trello* [Code]

Nascenia Ltd. — Dhaka, Bangladesh

Software Engineer, Intern — Nov, 2019 - Jan, 2020

Worked on a large-scale government project on VAT automation, **Project Uddogi**
This project involved — *Ruby on Rails, HTML, CSS, Heroku* [Article]

Teaching Experience

MasterCourse — Dhaka, Bangladesh

Data Science Instructor — Sep, 2022 - Present

Dokkho Data Science Career Program — [Course URL]

The goal of this program is the following to bridge the gap between academia and industry in the field of Data Science and Machine Learning and build the next generation of data scientists and machine learning engineers.

Amar iSchool — Dhaka, Bangladesh

Head Course Instructor — Mar, 2021 - Dec, 2021

Applied Data Science and Machine Learning — [Course Repository]

Our prime goal is to empower the local ML enthusiasts with a limited background CS and Mathematics.
- Taught data science and machine learning for production
- Gave 10 lectures, all self-developed
- Developed course assignments and projects

Skills

Programming Language: Python, C, C++, Ruby, Octave
Framework: Pytorch, fastai, Flask, Tensorflow, Keras, Ruby on Rails
Tools and Libraries: Git, LaTeX, Google Colab, HuggingFace, Google Cloud Platform
Languages: English *(fluent)*, Bangla *(fluent)*, Hindi *(medium)*, Urdu *(medium)*, Arabic *(beginner)*

Selected Awards & Achievements

An exhaustive list can be found here

– Won the OIC Scholarship that waived 75% of my undergraduate education costs (2% acceptance rate)
– **Champion**, 4^{th} Nurse Care Activity Recognition Challenge 2022. Won 100,000 JPY. **[P5]**
– **Champion**, 2^{nd} Nurse Care Activity Recognition Challenge 2020. Won a trip to Japan. **[P1]**
– **Finalist**, AI For Bangla Challenge 2021 (*among 147 teams*) **[P3]**

Extra-curricular Activities

– Travelled across some major locations in Bangladesh, Saudi Arabia and USA.
– **Contest Judge**, 16th December Programming Contest 2019
– **Former Programming Instructor**, IUT Programming Community
– **Former Vice President**, IUT Computer Society

References

Dr. Kevin Lybarger
Assistant Professor, Department of IST, George Mason University, VA, USA
Email: klybarge@gmu.edu

Dr. Marcos Zampieri
Assistant Professor, Department of IST, George Mason University, VA, USA
Email: mzampier@gmu.edu

Dr. Md. Ruhul Amin
Assistant Professor, Department of CIS, Fordham University, NY, USA
Email: mamin17@fordham.edu