

# Mohammad Sabik Irbaz

GitHub: [github.com/msi1427](https://github.com/msi1427)  
Portfolio: [msi1427.github.io/](https://msi1427.github.io/)

Email: [mirbaz@gmu.edu](mailto:mirbaz@gmu.edu)  
Phone: +17034593856

I am a Ph.D. student at George Mason University building patient-centered clinical NLP systems, advised by Dr. Kevin Lybarger. My current work develops retrieval-augmented generation (RAG) pipelines that map patient-authored narratives to standardized medical terminologies (RxNorm, SNOMED CT, CPT-4), combining hybrid retrieval, cross-encoder reranking, and LLM-based selection on multi-GPU HPC infrastructure. Before my Ph.D., I spent **2.5 years** in industry as a Data Scientist at Leadbook and Lead Machine Learning Engineer at Omdena, delivering end-to-end ML systems from data pipelines to deployment. My research spans **clinical concept normalization, text style transfer, RAG, reinforcement learning, and conversational agents**, with published work in the **Journal of Biomedical Informatics (JBI)** and a focus on bridging the linguistic gap between how patients describe their health and how clinical systems encode it.

## Research Interests

I work at the intersection of Natural Language Processing (NLP), Large Language Models (LLMs), and Clinical Machine Learning (ML), with an emphasis on systems that are both rigorous and deployable. My interests include, but are not limited to:

- i) **Clinical & healthcare NLP** — concept normalization, medical text simplification, and patient-facing language understanding [P8, P9, P10, P11, P12]
- ii) **RAG & LLM systems** — hybrid retrieval, reranking, and grounded generation for high-precision domains [P12]
- iii) **Text style transfer** — bridging lay and expert medical registers [P6, P8, P12]
- iv) **AI for social good** — applying NLP to public health, violence prevention, and low-resource settings [P1, P2, P5, P7, P9]

## Education

**George Mason University** Fairfax, Virginia, USA  
Ph.D. in Information Technology (INFT) Aug, 2023 - May, 2028 (*expected*)  
**Concentration:** Machine Learning Engineering  
**Supervisor:** Dr. Kevin Lybarger

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

## Research Experience

{<sup>α</sup>primary contributor; <sup>β</sup>secondary contributor; <sup>γ</sup>supervisor}

**Lybarger Language Lab** Aug, 2023 - Present  
**Graduate Research Assistant**  
*Department of IST, George Mason University, Virginia, USA*  
**Advisor(s):** Dr. Kevin Lybarger

- (1) <sup>α</sup>Health Text Simplification and Novel Strategies for RL — *NLP, Data Scraping, LLMs, RLHF* [P9]
- (2) <sup>α</sup>OCR and Vision-LLMs for IE from Handwritten Violence Reports — *OCRs, Multimodal LLMs* [P10]
- (3) <sup>β</sup>Enhancing Depression Management Through AI-Driven Conversational Agents — *LLMs, RLHF* [P11]
- (4) <sup>α</sup>Medical Concept Normalization in Patient Narratives — *LLMs, RAG, Text Style Transfer* [P12]

**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) <sup>α</sup>COVID-19 Public Policy Conformity Analysis — *NLP, Data Extraction, Topic Modeling, NLI, EDA* [P7]

Department of CSE, IUT, Dhaka, Bangladesh

Advisor(s): Dr. Md. Kamrul Hasan, Dr. Hasan Mahmud

(1) <sup>β</sup> Automated wheelchairs for semi-paralyzed patients. — *Signal Processing, ML, Arduino*

(2) <sup>α</sup> Medical Expertise Text Style Transfer — *NLP, Active Learning, Text Style Transfer* [P6]

Reviewing Experience

- 1 journal paper for **Communications Medicine, Nature, 2025**
- 2 conference papers for **IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023**
- 2 conference papers for **Empirical Methods in Natural Language Processing (EMNLP), 2022**

Publications

More updated list [here](#) or [Google Scholar](#)  
{\*equal contribution, <sup>c</sup>corresponding author}

[P12] **Mohammad Sabik Irbaz**, Tanvir Rouf Shawon, Saadat Hasan Khan, Vladimir Cardenas, Farrokh Alemi, Kevin Lybarger<sup>c</sup>. **“Enhancing Medical Concept Normalization in Patient Narratives with Style Transfer, Retrieval-Augmented Generation, and Conversational Clarification”**. *funded by Patient-Centered Outcomes Research Institute (PCORI). [Work in Progress]*

[P11] Kevin Lybarger<sup>c</sup>, Özlem Uzuner, Giridhar Kaushik Ramachandran, Hadeel Elyazori, **Mohammad Sabik Irbaz**, A. H. M. Rezaul Karim. **“Natural Language Processing of Clinical Data”**. *Springer Handbook of Data Engineering. [Accepted for Publication in October 2026]*

[P10] Md Tanvir Rouf Shawon, **Mohammad Sabik Irbaz**, Vladimir Franzuela Cardenas, Hadeel R. A. Elyazori, Keerti Reddy Resapu, Yili Lin, Vladimir Franzuela Cardenas, K. Pierre Eklou, Farrokh Alemi, Kevin Lybarger<sup>c</sup>. **“A Patient Simulation Framework for Risk Assessment of Conversational Healthcare AI: Evaluation of an Antidepressant Decision Aid”**. *Selected for presenting to Goerge Mason University AI-in-Gov Council and funded by Patient-Centered Outcomes Research Institute (PCORI). [Under Review at Journal of Medical Internet Research (JMIR)]*

[P9] **Mohammad Sabik Irbaz**, Md. Mushfiquir Rahman, Katherine Scafide, David Lattanzi, Janusz Wojtusiak, Kevin Lybarger<sup>c</sup>. **“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. [Work in Progress]*

[P8] Md Mushfiquir Rahman\*, **Mohammad Sabik Irbaz\***, Kai North, Michelle S Williams, Marcos Zampieri, Kevin Lybarger<sup>c</sup>. **“Health Text Simplification: An Annotated Corpus for Digestive Cancer Education and Novel Strategies for Reinforcement Learning”**. *Journal of Biomedical Informatics, Volume 158, October 2024*

[P7] **Mohammad Sabik Irbaz**, Rupak Sarkar, Md. Ruhul Amin<sup>c</sup> 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<sup>c</sup>. **“c-MSD: A Curated Dataset for Medical Expertise Style Transfer using Active Learning Approach”**. *Under Review.*

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

[P4] Alif Ashrafee\*, Akib Mohammed Khan\*, **Mohammad Sabik Irbaz<sup>c</sup>**, 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<sup>c</sup>**, and MD Abdullah Al Nasim. **“End-to-End Natural Language Understanding Pipeline for Bangla Conversational Agents”**. *Proceedings of the 20<sup>th</sup> IEEE International Conference on Machine Learning and Applications, 2021* [Presentation]

[P2] **Mohammad Sabik Irbaz<sup>c</sup>**, 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, 2021. Springer, Singapore.*

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

Industry Experience

**Leadbook Pte. Ltd.**  
**Data Scientist**  
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.

Singapore  
Dec, 2021 - Feb, 2023

**Omdena Inc.**  
**Lead Machine Learning Engineer**  
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]

New York, USA  
Apr, 2021 - Jun, 2021

**Pioneer Alpha Ltd.**  
**Lead Machine Learning Engineer**  
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]

Dhaka, Bangladesh  
Jan, 2021 - Oct, 2021

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

Dhaka, Bangladesh  
Nov, 2019 - Jan, 2020

Teaching Experience

**MasterCourse**  
**Data Science Instructor**  
**Dokkho Data Science Career Program** — [Course URL]  
The goal of this program was 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.

Dhaka, Bangladesh  
Sep, 2022 - May, 2026

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++, Javascript, Java, Ruby, Octave, Matlab  
**Framework:** Pytorch, fastai, Flask, LangChain, LangGraph, Tensorflow, Keras, Ruby on Rails  
**Tools and Libraries:** Git, LaTeX, Google Colab, HuggingFace, Google Cloud Platform, AWS, Azure  
**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<sup>th</sup> Nurse Care Activity Recognition Challenge 2022. Won 100,000 JPY. **[P5]**
  - **Champion**, 2<sup>nd</sup> Nurse Care Activity Recognition Challenge 2020. Won a trip to Japan. **[P1]**
  - **Finalist**, AI For Bangla Challenge 2021 (*among 147 teams*) **[P3]**
  - **Runners Up**, IUT Computer Programming Contest 2019 (*among 77 teams*)

References

---

**Dr. Kevin Lybarger**

Assistant Professor, Department of IST, George Mason University (GMU), VA, USA  
Email: klybarger@gmu.edu

**Dr. Farrokh Alemi**

Professor, Department of Health Administration, Policy, and Informatics (HAP), GMU, VA, USA  
Email: fa3277@gmu.edu

**Dr. Özlem Uzuner**

Professor and Department Chair, Department of IST, GMU, VA, USA  
Email: ouzuner@gmu.edu