Masum Hasan

PhD Student, Computer Science,
University of Rochester, NY
m.hasan@rochester.edu | (585) 360 3726
Linked in: masum6 | GitHub: Masum06 | Skype: youngladesh@live.com
https://masumhasan.net

Research Interest

Natural Language Processing, Human Computer Interaction, Prompt Engineering for large LM's, Low Resource NLP, Machine Learning for Source Code: Summarization/Synthesis/Search/Repair, Machine Learning, Artificial Intelligence.

Achievements

Best Student Paper Award (2018):
- International Conference on Bangla Speech and Language Processing – 2018
- Paper: *"Recognition of Bengali Handwritten Digits Using Convolutional Neural Network Architectures"*

Kaggle Contest Winner (2018):
- Bengali Handwritten Digit Recognition Competition
- https://www.kaggle.com/c/numta/leaderboard
- 99.3559% accuracy in NumtaDB, the largest dataset of Bengali Handwritten Digits

2nd Runner up at Robi Datathon (2019):
- First Data Science Hackathon in Bangladesh

NSysS Software Project Award (2016):
- http://floaterfb.blogspot.com/

BUET Admission Test Merit Scholarship
- For placing 20th in BUET admission test

Accepted Papers

- CoDesc: A Large Code–Description Parallel Dataset. **Masum Hasan**, Tanveer Muttaqeen, Ishtiaq Niloy, Kazi Mehrab, Tahmid Hasan, Mahim Pantho, Wasi Ahmad, Rifat Shahriyar, Anindya Iqbal
  Venue: **ACL Findings 2021**
- Using a Balanced Scorecard to Identify Opportunities to Improve Code Review Effectiveness: An Industrial Experience Report. **Masum Hasan**, Anindya Iqbal, Amiangshu Bosu, Mohammad Rafid Ul Islam, A.J.M. Imtiajur Rahman.
  Venue: **EMSE** (Journal), *Impact Factor: 8.41*
- Review4Repair: Code Review Aided Automatic Program Repairing. Faria Huq, **Masum Hasan**, Mahim Pantho, Sazan Mahbub, Anindya Iqbal, Toufique Ahmed.
  Venue: **IST** (Journal), *Impact Factor: 6.93*
- Hitting your MARQ: Multimodal ARgument Quality Assessment. Md Kamrul Hasan, James Spann, **Masum Hasan** Md. Saiful Islam, Kurtis Haut, Rada Mihalcea and Ehsan Hoque.
  Venue: **EMNLP 2021**
- Not Low-Resource Anymore: Aligner Ensembling, Batch Filtering, and New Datasets for Bengali-English Machine Translation.Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, **Masum Hasan**, Madhusudan Basak, M. Sohel Rahman, Rifat Shahriyar.
  Venue: **EMNLP 2020**
- Recognition of Bengali Handwritten Digits Using Convolutional Neural Network Architectures. **Md Mahmudul Hasan**, Md Rafid Ul Islam, Md Tareq Mahmood.
  Venue: **ICBSLP 2018**
- Early Detection of Earthquake Using Satellite Based Quantum Computing.Akhter Al Amin, **Mahmudul Hasan**, Kazi Sinthia Kabir, Tanzila Choudhury, ABM Alim Al Islam.
  Venue: **ICCSNT, 2015**

On Archive!

- Text2App: A Framework for Creating Android Apps from Text Descriptions. **Masum Hasan**, Kazi Sajeed Mehrab, Wasi Uddin Ahmad, Rifat Shahriyar.
  Project page: https://text2app.github.io/
  Venue: **EMLNLP 2022**
- Auto-Gait: Automatic Ataxia Risk Assessment with Computer Vision on Gait Task Videos. Wasifur Rahman, **Masum Hasan**, Md Saiful Islam, Titilayo Olubajo, Jeet Thaker, Abdelrahman Abdelkader, Phillip Yang, Tetsuo Ashizawa, Ehsan Hoque.
  Venue: **ML for Healthcare 2022**
- BERT2Code: Can Pretrained Language Models be Leveraged for Code Search? Abdullah Al Ishtiaq, **Masum Hasan**, Md. Mahim Anjum Haque, Kazi Sajeed Mehrab, Tanveer Muttaqueen, Tahmid Hasan, Anindya Iqbal, Rifat Shahriyar

On-going Projects

- **A GPT-3 based Chatbot to prepare doctors for end-of-life communication**
  - ROC-HCI, URMC
- **Post-COVID19 PTSD Identification from Tweets**
  - Georgia Tech, UR Computer Science
- **Low Resource Neural Machine Translation Indigenous Languages in Bangladesh**
  - Beyond the Hills, URCS, Fordham U, BUET
- **DialoGPT for Bengali**
  - Bangladesh University of Engineering and Technology (BUET)

Past Research Projects

Sub-Tree Compression: A Novel Compression Algorithm for Source Code Data*‡ [code]

- Compress repetitive code structures to reduce source code data length.
- Lossless reduction of number of code tokens by 10~20%.
- 10% faster training for state-of-the-art code summarization network with no performance reduction

GPT-Bengali*‡

- Crawled and pre-processed 18GB Bengali Corpus
- Trained a GPT2 small Language Model (110M parameters) in Bengali
- New benchmark datasets and results for Bengali Language models

What You See Is What You Learn: Mitigating Gender Bias in Natural Language Data by Encrypted Gender Training*‡ [draft]

- Gender debiasing LM's in 4 steps: Gender Encryption, Training, Generation, Gender Decryption
- Technologies: NER, Char-LSTM, Coreference Resolution, GPT2 Language Model

* Project Lead ‡ Research Initiator

Professional Experience

Graduate Research Assistant, ROC-HCI, University of Rochester, NY *Aug 2021 to Present*

- Supervised by Dr. Ehsan Hoque
- Parkinson's, PTSD, analysis with Machine Learning

Research Assistant (RA), Applied Machine Learning Lab, BUET *Nov 2018 to Jul 2021*

- Supervised by Dr. Rifat Shahriyar, Dr. Anindya Iqbal
- ML for source code: Source Code Summarization/Synthesis/Search/Repair/Compression
- NLP for Bengali: GPT2, BERT, Bengali-English NMT

Deep Learning Research Intern, Gaze Technologies *Jun 2018 to Aug 2018*

- Object Detection, License Plate Recognition
- Tensorflow Object Detection API

R&D Intern, REVE Systems *Mar 2018 to Jun 2018*
- Optical Character Recognition
- Industry Standard Coding Practices

Cofounder and CEO, Bizzy Ltd. *Dec 2016 to Jul 2017*
- BizzyBd started as an easy-to-use website building tool for non-programmer professionals.

Education

PhD Student, Computer Science (CS)
University of Rochester (UR) ~ Started Aug 2021
Courses:
- CSC 444: Knowledge Representation and AI
- CSC 460: Technology and Climate Change
- CSC 400: Introduction to Research
- CSC 480: Computational Modelling and Limitations
- CSC 463: Data Management Systems

BSc. In Computer Science and Engineering (CSE)
Bangladesh University of Engineering and Technology (BUET) ~ Oct 2018
- Thesis: Mythbusting Gender Stereotypes Using Natural Language Processing

Standardized Tests

Graduate Record Examinations (GRE)
- Analytical Writing: 4.0, Verbal Reasoning: **154**, Quantitative Reasoning: **166**. **Total: 320**

Test of English as a Foreign Language (TOEFL)
- Reading: **28**, Listening: **29**, Speaking: **24**, Writing: **24**. **Total: 105**

Technical Skills

Programming Languages: Python, Java, C, C++, JavaScript, MATLAB, HTML, SQL, MySQL, LaTeX, Arduino, AVR Programming.

Library/Framework/Others: Numpy, **Keras**, **PyTorch**, Tensorflow Object Detection API, Django, **OpenNMT**, **HuggingFace**, JQuery, Google SyntaxNet, Bootstrap.

Online Courses

- Machine Learning, Coursera.
- Deep Learning Specialization, Deeplearning.ai
- Practical Deep Learning for Coders, Fast.ai
- [Book] Neural Network and Deep Learning, Michael A. Nielsen
- [Partial] CS224n: Natural Language Processing with Deep Learning, Stanford University
- Essence of Linear Algebra, 3blue1brown

Blogs and Articles

Machine Learning:
- Absolute Beginner's Guide to Machine Learning and Deep Learning
- Setup and Run fast.ai in Amazon AWS

Graduate Admission:
- How I got 4 Ph.D. offers in the US with a CGPA 2.79 — and what you can learn from it

Extracurricular Activities

- Created an introductory Machine Learning course in my native language Bengali that dives into theory, practice, and intuition behind neural networks.
- Mentor at Camera to Chess project, Google Developer Student Club, University of Rochester
- Leisure: reading, rock climbing, running, chibi drawing, existential dread