

Sumon Kanti Dey

<https://sumonkantidey.github.io>
github.com/sumonkantidey | kaggle.com/sumon23
linkedin.com/in/sumon-kanti-dey
sumonkantidey23@gmail.com | +8801879862842

EDUCATION

EMORY UNIVERSITY | PHD IN COMPUTER SCIENCE AND INFORMATICS

Aug 2024 - Pres | Atlanta, Georgia

NOAKHALI SCIENCE AND TECHNOLOGY UNIVERSITY | B.Sc. IN COMPUTER SCIENCE

Mar 2014 - Dec 2018 | Noakhali, Bangladesh • GPA: 3.32 / 4.00

EXPERIENCE

EVCO.GLOBAL | DATA ENGINEER

Mar 2023 - Jun 2024 | Singapore

- Developing data pipelines to collect, clean, and analyze data from various sources (APIs, web scraping, client files).
- Designing and implementing scalable analytical solutions for electric vehicle charging infrastructure to accurately assess and predict battery capacity and 'State of Health'.
- Using Business Intelligence (BI) tools like Tableau and Plotly Dash to build a custom data dashboard, improve decision-making, and drive business success.

LEADBOOK.COM | DATA SCIENTIST

Apr 2020 - Feb 2023 | Singapore

- Worked on refactoring data structures for indexing and searching big data using Spark and Elastic Search.
- Maintain a data pipeline to ingest data from various sources into the data warehouse correctly, efficiently, and securely using Kafka, Redis Queue, and SQL.
- Implemented an email pattern generation system based on 180+ million data; verified and tracked them using a 30-node postfix cluster resulting in an overall accuracy of 87%.
- Used Python and Flask to work on 5+ scalable microservices and backend APIs. Worked with containers as well, including Docker and Kubernetes.
- Developed and deployed transformer-based pre-trained language models (PLMs) to classify companies into 30+ different industries and to estimate employee count and revenue, achieving a remarkable 3x faster model inference speed using ONNX.
- Designed a heuristic-based information retrieval tool to extract a company's name and postal address from its website.

HIFI DIGITAL LTD | PYTHON DEVELOPER

Mar 2019 - Mar 2020 | Banani, Dhaka

- Implemented large-scale information retrieval tool using Selenium and BeautifulSoup.
- Built an algorithm to parse table data from the attached email pdf or the email body.
- Developed a real-time road condition monitoring system using a pre-trained Faster R-CNN InceptionV2 model. Used the Django Rest Framework and SQLite database for the backend.
- Handled end-to-end development including designing, coding, and testing during each sprint cycle.

RESEARCH INTERN | VIS4SENSE RESEARCH GROUP

Jan 2019 - Mar 2019 | Middlesex University, London, UK

- I have conducted research under the supervision of Prof. Kai Xu in the areas of Data Science, Visual Analytics, and HCI.
- Applied ML techniques to break the user browsing history data collected by the SenseMap tool into smaller, meaningful chunks to understand users' sense-making strategies.
- Personalized the browsing experience of users by categorizing their online presence and analyzing their interests and goals, contributing to the development of Web 3.0.

TEACHING ASSISTANT | NOAKHALI S&T UNIVERSITY

Jun 2017 - Jun 2018 | Noakhali, Dhaka

- CSTE 2202 Data Structure and Algorithms Lab (2017 & 2018 session).
- Designed Course material, lesson plans, and homework assignments.

TECHNICAL SKILLS

Data science libraries : NumPy, Pandas, scikit-learn, Matplotlib, Tweepy, Keras, PyTorch.

Framework/Workflow management : Flask, Django, FastAPI, Apache Airflow.

Programming languages : Python, C/C++, Java, JavaScript, D3.js.

Cloud platforms : Amazon AWS, Google Cloud, DigitalOcean.

XAI toolkits : SHAP, LIME, ExplainerDashboard, ELI5.

Databases : Oracle, MySQL, PostgreSQL, MongoDB.

Containerization technologies : Docker, Kubernetes.

Data visualizations : Tableau, Plotly, Streamlit.

CI/CD, version control : Terraform, Git.

Big data stack : Spark, Hadoop.

Data streaming : Apache Kafka.

Tools : Bash, Linux, Vim.

PROJECTS

COVID19 FAKE NEWS DETECTION IN ENGLISH | [GITHUB](#)

Oct 2020 - Dec 2020

- Constraint@AAAI2021 shared task where the target was to predict whether the committed tweet was real or fake.
- Used transformers-based LMs variant and utilized ensemble-based approach based on the top three models.
- Our best model achieved the weighted F1-score of 0.978 on the test set.

JIGSAW MULTILINGUAL TOXIC COMMENT CLASSIFICATION | [GITHUB](#)

Mar 2020 - Jun 2020

- The target of this competition is to identify toxicity in online conversations.
- Used Multi-Stage training based on different architectures on top of XLM-RoBERTa.
- The small-scale model finished at 151st place out of 1621 teams.

BENGALI.AI HANDWRITTEN GRAPHEME CLASSIFICATION | [GITHUB](#)

Dec 2019 - Mar 2020

- Multi-class classification challenge where need to classify three constituent elements from an image: grapheme root, vowel diacritics, and consonant diacritics.
- Applied active Contour Based Segmentation for image preprocessing. 64x64 pixels with one single channel image fit into a MultiOutput CNN model to classify three classes from the image.
- Build a Flask WebApp that will help to identify real-time Grapheme with three constituent elements.

QUORA INSINCERE QUESTIONS CLASSIFICATION | [GITHUB](#)

Oct 2019 - Oct 2019

- The objective of this project is to predict whether a question asked on Quora is sincere or not.
- Applied 5-fold stack attention-based Bi-GRU with different embedding (Glove, Word2Vec, FastText).
- Developed a web app for real-time question classification into sincere and insincere categories.

APTOS 2019 DIABETIC RETINOPATHY DETECTION TO STOP BLINDNESS | [KAGGLE](#)

Jun 2019 - Sep 2019

- An extensive set of retina images was taken using fundus photography under various imaging conditions. We need to rate each image for the severity of diabetic retinopathy on a scale of 0 to 4.
- Used an ensemble of three different CNN architectures, including Efficientnet-B7, Inception-v4, and SE-ResNeXt50, to report the final prediction.
- Secured 199th Position (top 7 percent) among 2931 teams.

RNN AND CNN NETWORKS FOR SARCASM DETECTION VIA AUXILIARY AND EMOTION FUSION | [BACHELOR THESIS](#)

May 2018 - Nov 2018

- Collected 32k tweets with sarcastic and non-sarcastic markers and developed an annotated sarcasm dataset.
- Developed an improved context, contrast, and semantic modeling network for sarcasm detection.
- Proposed DNNs architectures followed by additional handcrafted features and emotional components to detect sarcasm from Tweet.

SELECTED PUBLICATIONS

1. Md Rashad Al Hasan Rony, Sudipto Kumar Shaha, Rakib Al Hasan Joy, **Sumon Kanti Dey**, amzad Hossain rafi, Ashraf Hasan Sirajee and Jens Lehmann: [BanglaQuAD: A Bengali Open-domain Question Answering Dataset](#). LREC-COLING 2024
2. Md. Rezaul Karim, **Sumon Kanti Dey**, Tanhim Islam, Md. Shajalal, Bharathi Raja Chakravarthi: [Multimodal Hate Speech Detection from Bengali Memes and Texts](#). In international conference on Speech & Language Technology for Low-resource Languages (SPELLL,2022)
3. Md. Rezaul Karim, **Sumon Kanti Dey**, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Bharathi Raja Chakravarthi, Md. Azam Hossain, Stefan Decker: [DeepHateExplainer: Explainable Hate Speech Detection for Under-resourced Bengali Language](#). In IEEE International Conference on Data Science and Advanced Analytics (DSAA’2021)

ACHIEVEMENTS

- 4th position among 158 teams in **AI for Bangla 2.0** competition, hosted by Bangladesh Computer Council of ICT Division, Bangladesh. 2023
- Kaggle **Competition Expert** (Ranked 3471 among 144,292 Kagglers worldwide). 2020
- 24th place at SUB Inter-University Programming Contest. 2017
- 17th place at DIU ACM ICPC Dhaka Region Worlds finals Warm-up. 2016
- 1st place at CSTE carnival’15 Programming Contest,NSTU. 2015
- Top 1% in Noakhali S&T University Entrance Exam. Ranked 54 among thousands participants. 2014

EXTRA CURRICULAR ACTIVITIES

- General Secretary at the departmental computing club, contributing to the growth of a community of 150+ people. 2017-2018
- Organized 8+ programming competitions, 1 boot-camp, 6+ workshops, and 10+ seminars. 2015-2018
- Served as an active member of the NSTU Blood Donors Society. 2015-2018
- Chief Judge at Intra NSTU PI Day Programming Contest(26 teams). 2018
- Organizer and Mentor at Junior Programming Boot-camp *goicpc18* [45 participants]. 2017
- Workshop Trainer at CSTE Club Competitive Programming. 2016-2017
- Judge and Problemsetter at NSTU Intra CSTE programming Contest [30 teams]. 2017