Md. Motahar Mahtab

+8801869977076 | md.motahar.mahtab@g.bracu.ac.bd | Linkedin | GitHub | Portfolio

Education

BRAC University Dhaka, Bangladesh
CGPA: 3.99; B.Sc in Computer Science & Engineering *May 2018 – May 2022*

Notredame College Dhaka, Bangladesh
GPA:5.0; HSC in Science *2015 – 2017*

Experience

ML Engineer (Remote) Oct. 2024 – Present
Delineate Inc. *Cambridge, MA, USA*

- Created two data extraction pipelines (Covariate and General) for Quantitative Systems Pharmacology (QSP) related research papers. These systems use multi-stage retrieval, context reranking and filtering approaches that improve their accuracy on QSP QA by 7.37% on average.
- To enhance RAG in aforementioned extraction systems, collaborated in improving an advanced paper layout system using MinerU. The improved layout system adds functionalities for breaking a figure into its subfigures and also extracting information like captions, footnotes, etc. It can also add respective legend images if they are not present in the subfigures and store them. This additional processing increases IR performance from images/plots. The chunks for texts, images and tables also contain location information like which section/subsection of the data they are from and metadata like caption, summary and footnote which improves RAG accuracy by 12.36% on average.
- Collaborated on creating a bounding box matching algorithms that can match a user cropped figure to the correct subfigure in the paper.

Jr. ML Engineer Sep. 2022 – Sep. 2024
Giga Tech Ltd. *Gulshan, Dhaka, Bangladesh*

- Created new state-of-the-art systems for a plethora of Bangla NLP tasks e.g. Named Entity Recognition (NER), Parts of Speech (POS), Lemmatization, Question Answering, and Emotion recognition. Performed R&D on increasing performance beyond the current state-of-the-art to achieve 90% KPI on ML modules. Two such systems Bangla Lemmatization and Emotion recognition are publicly available at https://github.com/eblict-gigatech/BanLemma and https://sentiment.bangla.gov.bd respectively.
- The NER and PoS classification module establishes new state-of-the-art results on Bangla NER and PoS datasets by a hierarchical majority voting mechanism among external contexts retrieved from a Knowledge Base.
- The Question Answering (QA) module establishes new state-of-the-art results on Bangla datasets including SQuAD-bn (translated from the SQuAD-2.0 and TyDI-QA English QA datasets) by a modified loss function to balance performance among null and non-null questions.
- The Bangla Lemmatization system lemmatizes words based on their parts of speech class within a given sentence. It achieves an accuracy of 96.36% when tested against a manually annotated test dataset by trained linguists and demonstrates competitive performance in three previously published Bangla lemmatization datasets.
- Created GPT4o inference pipeline for Bangla NER and Coreference Resolution systems using ReAct prompting method achieving comparable performance against finetuned systems.
- Created pipeline for Natural Language generation (NLG) in Bangla for both encoder models like BERT and auto-regressive models like GPT2. Analyzed and overcame common issues like repetitive text generation, and unmeaningful word generation in NLG for Bangla.
- Created data augmentation pipeline to handle the class imbalance problem in sequence tagging tasks. The augmentation pipeline followed a sentence resampler method that selects sentences that contain rare classes/tags. Experimented with different sorts of loss functions like Dice, Focal and CurricularFace loss to handle Data imbalance problems.
- Optimized deployment of LLMs using Optimum (for ONNX conversion) and Nvidia TensorRT(TRT) format for further optimization. Used PyTorch Profiler to identify inference bottlenecks. Used Nvidia Triton Inference Server (TIS) as the default ML inference server for concurrent request serving and scheduling, batch inference and response caching in MongoDB. Used Locust for load testing and pytorch profiler to reduce bottlenecks.
- Created REST APIs using FastAPI for hosting ML inference endpoints.
- Used Qdrant vector DB for fast semantic searching, Dask to analyze and query big dataframes, DVC for dataset versioning and MLflow for model, artifact and experiment versioning.

PUBLICATIONS

BanNERD: Context-Driven Approach for Bangla Named Entity Recognition 2024
2025 Conference of the Nations of the Americas Chapter of ACL (NAACL) *New Mexico*

- A Bangla Named Entity recognition method named BanNERCEM (Bangla NER context-ensemble method) which outperforms existing approaches on Bangla NER datasets and performs competitively on English datasets using lightweight Bangla pretrained LLMs. Our approach passes each context separately to the model instead of previous concatenation-based approaches and performs an entity-aware majority voting among the contexts' predictions. It achieves the highest average macro F1 score of 81.85% across 10 NER classes, outperforming previous approaches and ensuring better context utilization.
- The most extensive human-annotated and validated Bangla Named Entity Recognition Dataset to date named BanNERD, comprising over 85,000 sentences was introduced. To ensure the dataset's quality, expert linguists developed a detailed annotation guideline tailored to the Bangla language. All annotations underwent rigorous validation by a team of validators, with final labels being determined via majority voting, thereby ensuring the highest annotation quality and a high IAA score of 0.88. In a cross-dataset evaluation, models trained on BanNERD consistently outperformed those trained on four existing Bangla NER datasets.
- Meta review score: 4.0/5.0

BanLemma: A Word Formation Dependent Rule Based Lemmatizer 2023
The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP); H-Index:176 *Singapore*

- The Bangla Lemmatization system lemmatizes words based on their parts of speech class within a given sentence. Unlike previous rule-based approaches, this system analyzes the suffix marker occurrence according to the morpho-syntactic values and then utilizes sequences of suffix markers instead of entire suffixes. To develop the rules for suffixes of each Parts of Speech and stripping approach, a large corpus of 90.65M unique Bangla sentences from various domains were analyzed to observe the word-formation of inflected words. The lemmatizer achieves an accuracy of 96.36% when tested against a manually annotated test dataset by trained linguists and demonstrates competitive performance in three previously published Bangla lemmatization datasets.
- Source: https://github.com/ eblict-gigatech/BanLemma

BanglaBait: Semi-Supervised Adversarial Approach for Clickbait Detection 2023
International Conference on Recent Advances in Natural Language Processing; H-Index:36 *Varna, Bulgaria*

- First Bangla Clickbait News Article Dataset containing 15,056 data instances. Investigated various semi-supervised learning methods and compared them with supervised learning methods to prove the former's superiority.
- Source: https://github.com/mdmotaharmahtab/BanglaBait

GAN-BERT Approach for Bengali Text Classification with Few Labeled Examples 2022
International Conference on Distributed Computing and Artificial Intelligence (DCAI); H-Index:36 *Portugal*

- Trained state-of-the-art Transformer networks in adversarial fashion using Generative Adversarial Network (GAN) to achieve superior performance when labelled dataset size is too small. First Bangla Paper to investigate the application of GAN-BERT on Bangla text classification tasks.
- Source: https://link.springer.com/chapter/10.1007/978-3-031-20859-1

Technical Skills

ML Libraries: PyTorch, PyTorch Lightning, Huggingface, LangChain, LangGraph, AutoGPT, OpenCV, Flair, OpenNMT, AllenNLP, NLTK, Pandas, Matplotlib, NumPy
Web Frameworks: Flask, Django, FastAPI, Streamlit
Developer Tools: Git, Docker, Locust, pre-commit
DBMS: MongoDB, PostgreSQL, Supabase, SQLAlchemy. **ML Tools**: Triton, Dask, DVC, MLflow, Elasticsearch, Qdrant, Ray, Wandb, TensorBoard, Pytorch Profiler
Programming: Python, Bash, SQL

Awards

BRAC University Intra University Programming Contest | *Winner* 2019
BRAC University Merit Scholarship Award 2018 – 2022
Dean's Prestigious List Award 2022

Projects

Bangla Clickbait Detector App | *Pytorch, Streamlit, Node.js* 2022

- Demo app created as a part of research work on Bangla Clickbait Detection using GAN-Transformers. It takes a Bangla article title as input and outputs whether the title is a clickbait or non-clickbait along with the prediction probability score. GAN-Transformers is a Transformer network trained in a generative adversarial training framework.
- Project Link: https://github.com/mdmotaharmahtab/Bangla-Clickbait-Detector-App

Bangla Article Headline Categorizer App | *Pytorch, Streamlit, Node.js* 2021

- Can categorize Bangla article headlines into eight different categories - Economy, Education, Entertainment, Politics, International, Sports, National, and Science & Technology
- Models used: State-of-the-art Bangla ELECTRA model, Dataset used: Patrika Dataset. - contains $400k$ Bangla news articles from prominent Bangla news sites.
- Project Link: https://github.com/mdmotaharmahtab/Bangla-Headline-Categorizer-App

EBRAC - Online Learning App | *Django, Bootstrap, Node.js* 2020

- A comprehensive online education platform where instructors can create different courses, upload course content, enrol students, see students' marks, prepare questions, take quizzes etc.
- Students can enrol in courses, view course contents, participate in exams and see results.
- Project Link: https://github.com/mdmotaharmahtab/EBRAC

Veggie: Vegetarian Recipe Maker App | *Django, Bootstrap, Node.js* 2019

- This web app allows users to view different vegetarian recipes, and see their total calories, nutrients like protein, carbohydrate, fat and their ingredients.
- Users can create their own vegetarian recipes by mixing different ingredients available on the web app. They can also see the total nutrients and calories of their created recipe
- Project Link: https://github.com/mdmotaharmahtab/veggie

Certifications

AWS Machine Learning Foundations | *Udacity* 2022

- Learned how to prepare, build, train, and deploy high-quality machine learning (ML) models with Amazon SageMaker and use AWS AI Services (i.e. AWS DeepLens, AWS DeepRacer, and AWS DeepComposer).
- Certificate: http://tinyurl.com/motaharudemycertificate

Articles

Medium

- Sparse Transformers Explained — URL: medium.com/@mahtab27672767/sparse-transformers-explained-part-1-aacbe10dca4a

Open Source Contributions

- https://github.com/flairNLP/flair/pull/3449 ; **Flair** is a framework for state-of-the-art NLP embeddings and training sequence models. Contributed to fixing a bug in the Flair framework which was causing incorrect prediction distribution output for a sequence of tokens in sequence classification tasks (Chosen to be merged in their next version release.)

References

Dr. Ruhul Amin Shajib

- Assistant Professor, Dept. of Computer & Information Science, Fordham University, Bronx, NY 10458, United States Email: shajib.sustgmail.com

Dr. Nabeel Mohammed

- Associate Professor, Dept. of Electrical & Computer Engineering, North South University, Dhaka, Bangladesh, Email: nabeel.mohammednorthsouth.edu

Dr. Farig Yousuf Sadeque

- Associate Professor, Dept. of Computer Science & Engineering, BRAC University, Dhaka, Bangladesh Email: farig.sadeque@bracu.ac.bd