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*

Experience

Jr. ML Engineer Sep. 2022 – Present
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, and Emotion recognition. Bangla Lemmatization and Emotion recognition systems are publicly available at github.com/eblict-gigatech/BanLemma and sentiment.bangla.gov.bd respectively.
- 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.
- Created REST APIs using FastAPI for hosting ML inference endpoints. Used MongoDB for response caching in NVIDIA Triton.
- 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.
- 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.

Research Assistant (remote) Sep 2021 – Dec 2021
Qatar Computing Research Institute *Dhaka, Bangladesh*

- Pretrained a HuBERT model on Bangla ASR dataset for joint task of speech and speaker recognition pipeline using SpeechBrain.
- Assisted in enriching existing open source Bangla ASR datasets by adding more scripted audio and correcting existing annotation.

PUBLICATIONS

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

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

- State-of-the-art Bangla Rule Based Lemmatizer which proposes a novel iterative suffix stripping approach based on the part-of-speech tag of a word.Shows superior performance than all previously published Bangla lemmatization methods on existing datasets.
- Source: https://github.com/ eblict-gigatech/BanLemma

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

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

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
ML Tools: Triton, Dask, DVC, MLflow, Elasticsearch, Qdrant, Ray Tune, Wandb, TensorBoard, Pytorch Profiler
Programming: Python, Bash, SQL

Awards

BRAC University Intra University Programming Contest | *Winner* 2019

BRAC University Merit Scholarship Award 2018 – 2022

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

Flair 2024

- 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.)
- https://github.com/flairNLP/flair/pull/3449