Abu Ubaida Akash

Curriculum Vitae (February 27, 2023)

Dhaka, Bangladesh

akash.ubaida@gmail.com abuubaida.github.io github.com/AbuUbaida +8801858204222

Research Interests

Natural Language Processing, Computational Linguistics, Machine Learning, Information Retrieval

Academic Credentials

Ahsanullah University of Science and Technology (AUST) Dhaka, Bangladesh

Bachelor of Science in Computer Science and Engineering, CGPA: 3.52/4.0 (Top 15%) **Apr. 2017 – Jan. 2022**

- **Major CGPA:** 3.72/4.0 [Last two years]
- **Major Courses:** Artificial Intelligence, Soft Computing, Pattern Recognition, Digital Image Processing, Computer Graphics, Distributed Database Systems, Formal Languages & Compilers, Operating System.
- **Thesis:** Development of Machine Learning Models for Crime Prediction using Historical Data [PDF]
  Supervisor: Prof. Dr. Mohammad Shafiul Alam [profile]

Publications

- **Abu Ubaida Akash**, Mir Tafseer Nayeem, Faisal Tareque Shohan, Tanvir Islam,
  Shironaam: Bengali Headline Generation using Auxiliary Information,
  In Proceedings of the European Chapter of the Association for Computational Linguistics: EACL 2023.
  [accepted] [PDF] [code]
- Faisal Tareque Shohan, **Abu Ubaida Akash**, Dr. Muhammad Ibrahim, Dr. Mohammad Shafiul Alam,
  Crime Prediction using Machine Learning with a Novel Crime Dataset,
  In Cybernetics and Systems Journal: UCBS. [under-review] [arXiv]
- **Abu Ubaida Akash**, Mir Tafseer Nayeem, Faisal Tareque Shohan, Samsul Islam,
  A Large-Scale Dataset, Criteria, and Models for Multilingual Headline Generation. [ongoing]

Professional Experience

Research Engineer (Speech & NLP) **Feb. 2023 – Present**

AIMS Lab, United International University Dhaka, Bangladesh

- Project: Automated and Intelligent Medical Scriber for Doctor Patient Conversation (AIMScribe)
- Designing protocols for collecting audio data and automatically transcribing them.
- Designing experiments for speaker diarization, NER, and summarization in the Bengali language.
- Applying different NLP, speech processing, and machine learning algorithms to produce deliverable output.
- Writing high-quality research articles on the AIMScribe project.

Machine Learning Research Engineer **Sep. 2021 – Apr. 2022**

Intelsense AI Dhaka, Bangladesh

- Implemented a G2P model for Bengali and gained state-of-the-art accuracy (99%) on unseen data.
- Prepared large-scale (nearly 600 hours) audio data for better Bengali ASR training.
- Speech synthesis: Implemented Coqui TTS models for low-resource languages like Bengali.
- Conversational AI: Developed AI-driven chatbots using Rasa Open Source.
- Bengali transcriber: Prepared the annotated corpus for the Bengali transcriber; already in use.

Machine Learning Research Intern **Jun. 2021 – Aug. 2021**

Intelsense AI Dhaka, Bangladesh

- Developed the pipeline for Bengali text normalization and punctuation restoration.
- Reviewed the literature on the related technologies.

Open-Source Contributions

BenSim

- Created a Python package for measuring the semantic similarity among sentences in the Bengali language.
- Measured similarities between BERT embeddings by applying Euclidean distance or Cosine similarity. [code]

mSentsTokenizer

- Developed a package in Python for tokenizing multilingual documents at the sentence level; currently supports 41 languages from 10 language families. [code]

Experimental Projects

- **Question Similarity Assessment using Transfer Learning:** Experimented the workflow of transfer learning from a general pre-trained language model (BERT) to a similarity measurement task. [code]
- **News Classification using Vanilla Transformer:** Implemented the transformer network from scratch and modified for text classification. [code]
- **B2E Neural Machine Translation with Seq2Seq Model using Attention:** In this experiment, Bengali to English machine translation was performed by training a sequence-to-sequence model with attention mechanism. GRU was incorporated in both the encoder and decoder with the attention layer. [code]
- **Character-Level Name Generation using LSTM:** Trained a language model using LSTM to generate English names character-by-character. Character level tokens were picked from the the conditional probability distribution using top-K sampler. [code]

Technical Skills

Languages: Python, C, C++, Java, Assembly 8086
Libraries: PyTorch, Tensorflow, Keras, HuggingFace, scikit-learn, Pandas, NumPy, Matplotlib, Seaborn
Back-end Engineering: Node.js, MongoDB, Express.js
Developer Tools: VS Code, Jupyter, GitHub, GCP
Miscellaneous: Git, Linux, LaTeX

Honors/Awards

- Robi-Datathon 2.0 Finalist (Top 6% among 358 Teams); organized by Robi Axiata Limited. (2022)
- Game Showcasing Competition 1st Runner-Up; organized by AUSTIDC. (Spring 2018)

Voluntary Services

- General Member at AUST Innovation and Design Club (AUSTIDC). (May 2017 – Dec. 2021)
- Communication Responsible at Mozilla Bangladesh. (Jan. 2018 – Jan. 2018)
- Content Developer at Durbin Labs Limited. (Jun. 2018 – Aug. 2018)

References

Dr. Mohammad Shafiul Alam
Professor and Former Head,
Department of CSE, AUST
Email: shafiul.cse@aust.edu

Arif Ahmad
Chairman & Head of Research
Intelsense AI Ltd.
Email: arif@intelsense.ai

Mir Tafseer Nayeem
Assistant Professor *(On Study Leave)*,
Department of CSE, AUST &
PhD Researcher, University of Alberta
Email: mnayeem@ualberta.ca