Samee Arif

samee.arif@lums.edu.pk • Website • Google Scholar • Github • Huggingface

RESEARCH INTEREST

My research focuses on Information & Communication Technologies for Development, Human-Computer Interaction, and Speech and Natural Language Processing. I am committed to leveraging these fields to create meaningful social impacts, particularly by enhancing accessibility and usability of technology.

EDUCATION

Bachelor of Science in Computer Science *Sep 2019 - May 2023*

Lahore University of Management Sciences (LUMS)

- Relevant Coursework: Artificial Intelligence, Machine Learning, Deep Learning, Natural Language Processing, Speech Processing, Principles and Techniques of Data Science, Computer Vision, Mathematical Foundations for Machine Learning and Data Science, Probability, Calculus II

RESEARCH EXPERIENCE

Student Counseling Chatbot | *Python, Next.js, React, FastAPI, Git, Google Colab* *Aug 2023 - Present*

- Developing a graduate assistant tool leveraging Large Language Models (LLMs) to provide educational counseling.
- Worked on prompt engineering for LLMs.
- Implemented multimodality by integrating Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) systems.

Urdu Question-Answering | *Python, Transformers, PyTorch, Google Colab* *Jan 2023 - Oct 2023*

- Developed a Question-Answering corpus for the Urdu language to address the limited resources available in the domain.
- Manually evaluated Seamless M4T and Google Translator for Urdu.
- Introduced EATS - a technique to preserve the answer spans in the translated context paragraphs and employed it to translate the SQuAD2.0 dataset to Urdu.
- Successfully generated 124,745 question-answer pairs and fine-tuned mBERT, XLM-RoBERTa, mT5 and LLaMA-2 on our dataset to achieve an **85.99% F1 Score** and **74.56% Exact Match**.
- First authored and published a research paper at **LREC-Coling 2024**.

Image-to-Speech Pipeline for Urdu Language | *Python* *Sep 2021 - Sep 2022*

- Evaluated Optical Character Recognition (OCR) models including Tesseract, EasyOCR, and Kraken on Nastaliq font.
- Established a pipeline to replicate scanned images using data augmentation to generate the dataset.
- Fine-tuned GANs to map the noisy images to clean images as a pre-processing module.
- Implemented a post-processing module based on BERT, Google search engine auto-correction and conditional random fields to enhance the model accuracy.
- Trained Tesseract to achieve a **1.53% Character Error Rate** and piped it with my Text-to-Speech (TTS) model.

WORK EXPERIENCE

Research Associate | *ActualAIz (LUMS)* *Aug 2023 - Present*

- Advisor(s): Dr. Agha Ali Raza, Dr. Ihsan Ayyub Qazi and Dr. Zafar Ayyub Qazi (LUMS).
- Working on developing a multimodal and multilingual graduate assistant tool leveraging large language models to provide educational counseling.

Research Assistant | *CSaLT (LUMS)* *Aug 2021 - May 2023*

- Advisor(s): Dr. Raza (LUMS), Dr. Awais Athar (EMBL-EBI).
- Worked on image-to-speech pipeline and Urdu question-answering system.

Teaching Assistant | *Machine Learning (LUMS)* *Fall 2022*

- Oversaw and facilitated learning for a cohort of 149 students. Designed and administered course quizzes, assignments and a project to gauge student understanding and progress.

Teaching Assistant | *Computational Problem Solving (LUMS)* *Fall 2021*

- Managed a 93-student cohort, designed quizzes, and labs and held weekly office hours.

PROJECTS

Speech Technologies | *Python, PyTorch, Transformers, React, FastAPI, Git* *Aug 2023 - Dec 2023*

- Fine-tuned Whisper and MMS ASR model, achieving a **13.01% WER**. Analyzed model quality and inference time, integrated quantization for faster inference, and utilized QLoRA for efficient fine-tuning.
- Trained MMS-TTS and YourTTS, adapting a VITS TTS framework script for training.
- Created a web-based audio annotation tool providing editable transcriptions and timestamps using ASR.

ConvoLense | *Python, Transformers FastAPI, Git* *Aug 2023 - Sep 2023*

- Evaluated speech-based (Wav2Vec2) and text-based (BERT, mT5, GPT, LLaMA) emotion classifiers.
- Used Bark to generate a synthetic conversation dataset between customer and customer service representative.
- Established a pipeline using my ASR model and LLM for emotion classification.

Arabic Handwriting Recognition | *Python* *Jan 2023 - May 2023*

- Applied transfer learning techniques to adapt the Urdu OCR model for recognizing handwritten Arabic in Naskh font.
- Utilized advanced pre-processing methods, such as skeletonization, to generate a synthetic handwritten dataset.

Image Captioning | *Python, PyTorch, Google Colab* *Jan 2023 - May 2023*

- Conducted an experimental fine-tuning of Swin-Transformer on the Indiana University - Chest X-Rays dataset, exploring its application in medical image analysis.

Fraudulent Job Prediction | *Python, Scikit-learn, Pandas Jupyter* *Sep 2022 - Dec 2022*

- Trained Logistic Regression, Support Vector Machine, and Random Forest classifiers to identify real versus fake job postings, achieving a **91% Accuracy**.
- Conducted comprehensive data cleaning and exploratory data analysis on the dataset.
- Authored and published an article on Medium detailing the project's methodology and outcomes.

Lane Analysis for Autonomous Vehicle | *Python, PyTorch, Google Colab* *Sep 2022 - Dec 2022*

- Created a lane-change warning system, integrating Lanenet for lane detection and YOLOv7 for vehicle detection.

Learning Management System | *MongoDB, Node.js, React, Git, Trello, Postman* *Jan 2022 - May 2022*

- Created a platform for schools to manage online education during the pandemic.

Speech-based Language Classifier | *Python, Scikit-learn, Jupyter* *Sep 2021 - Dec 2021*

- Recorded voice samples in English, Urdu, and a mix of both languages at 1600MHz.
- Developed and trained a neural network from scratch to classify speech using the recordings dataset.

FoodSwings | *HTML, CSS, Bootstrap, React, Node.js, MySQL, Postman* *Sep 2021 - Dec 2021*

- Implemented food delivery web application.

Neural Network from Scratch | *Python, NumPy, Jupyter* *Sep 2021 - Dec 2021*

- Developed a feed-forward neural network from scratch using NumPy and optimized it with Numba's JIT.

AWARDS

Dean's Honour List | *LUMS* *Fall 2020*
Dean's Honour List | *LUMS* *Spring 2019*

TECHNICAL SKILLS

Languages | *Python, C/C++, SQL, JavaScript, HTML/CSS*
Frameworks | *React, Node.js, Next.js, FastAPI*
Developer Tools | *Git, Docker, Google Cloud Platform, VS Code, Visual Studio*
Libraries | *pandas, NumPy, Matplotlib, TensorFlow, PyTorch, Keras, transformers, Streamlit*

PUBLICATIONS

[1] **Samee Arif**, Sualeha Farid, Awais Athar, and Agha Ali Raza, UQA: Corpus for Urdu Question Answering. In **LREC-COLING 2024** – Joint International Conference on Computational Linguistics, Language Resources and Evaluation. May 20-25, 2024, Torino (Italia). (**Ranks 6th from top in Computational Linguistics**)