YASH HATEKAR

+1 930-220-8410 ◇ yashhatekar1997@gmail.com ◇ Linkedin ◇ GitHub ◇ Portfolio

EDUCATION

Master's in Computer Science, Indiana University Bloomington - 3.9 CGPA 2022 - 2024
Bachelor's in Computer Science, Maharashtra Institute of Technology - 7.34 CGPA 2015 - 2019

SKILLS

Languages	Python, PySpark, SQL, R, Java, C++, C, Bash, HTML, Assembly, Hadoop, PLSQL, C#.
Libraries	PyTorch, Tensorflow, HuggingFace, Keras, FastText, Librosa, Whisper, LangChain, MLFlow, Prompt Engg, GenAI
Tools	MongoDB, Github, Tableau, Power BI, JIRA, Linux, Azure, Databricks, CI/CD, Google Coral, Raspberry pi, Arduino
Models	LLM's, BERT, CNN, LSTM, ARIMA, Statistical, Supervised and Unsupervised, Ensemble Learning, Deep Learning

EXPERIENCE

Data Science Fellow - Indiana University - *Bloomington, Indiana* May 2023 - May 2024

- Improved the efficiency of our speech-to-text processes which were used to extract interviewee dialogue from audio files across 13 different dialects of Spanish spanning 8 countries.
- Automated this process and reduced the manual annotation of data.
- Crafted an automated speech-to-text script and cross-analyzed transformers (BERT, SPANBERT, ROBERTA) and AI tools (Llama, Mistral, ChatGPT) using Ollama to perform speech-to-text and identify the importance of vowel spacing in Spanish utilizing Python, Tensorflow, and PyTorch.
- Reduced manual annotation of data by 80%.

Data Analyst - Deloitte - *Mumbai, India* Feb 2021 - Jul 2022

- Developed an insurance platform using Majesco Policy Administration System for Homesite. The platform was data-driven and required rigorous testing to ensure system stability and functionality.
- Developed and processed end-to-end testing for over 3 insurance products (Commercial Auto, BOP, and GL).
- Leveraged SQL to develop the testing processes, ensuring each product was thoroughly tested for potential issues.
- As a result, we were able to ensure the system's stability and functionality, contributing to the successful development of the platform.

Data Scientist - Deloitte - *Mumbai, India* Jan 2020 - Jan 2021

- Created a model to predict the number of deaths due to COVID-19.
- My task was to design a custom pandemic forecasting model that considered several factors including herd immunity, the number of hospital beds, and the severity of government restrictions.
- Designed a model stemming from SIR (Susceptible, Infected, Recovered) and ARIMA models, and utilized Power BI to summarize our findings in ten dashboards visualizing economic impacts on different states in the US.
- We were able to predict the number of deaths due to COVID-19 more accurately, providing valuable insights for decision-makers.

NLP Intern - Persistent Systems Ltd. - *Pune, India* Jul 2019 - Sep 2019

- As an NLP Intern at Persistent Systems Ltd., I was tasked with improving the feedback process for employees.
- Designed and deployed a True Feedback model using NLP tools to provide insight into areas of strength and opportunities for improvement for an employee.
- Utilized textual feature extraction tools like FastText, Spacy, Gensim, and Tensorflow to generate genuine feedback from clients for employees.
- We were able to provide more accurate and helpful feedback to the employees, aiding in their professional development.

Machine Learning Intern - Greenova Corporation - *Pune, India* Sep 2018 - May 2019

- Developing an Android application to detect the early onset of dementia.
- Integrated speech recognition and facial emotion recognition machine learning techniques into the application.
- Combined five speech features and seven facial emotions to develop the model using NLP, Android, and Firebase. The model incorporated elements such as a facial feature extraction model and a speech model.
- As a result, we were able to develop an Android application that could potentially detect the early onset of dementia, providing a valuable tool for early intervention.

PROJECTS

End to End Olympic Data Analytics

- Configured Azure data factory to ingest Paris 2024 Olympics data.
- Performed data transformation using Azure Databricks to transform raw data for future analytics.
- Designed and configured Azure Synapse Analytics to generate meaningful insights such as medal counts per country, athletes per country, and average number of entries by gender for each discipline.
- Created a dashboard to visualize the insights gathered from analytics using Power BI.

Bird Classification using Coral TPU

- Configured and developed a bird identification model on Google Coral TPU.
- Utilized Python and Tensorflow lite to implement classification model on the TPU.

AMWAL: Named Entity Recognition for Arabic Financial News

- Cross-analyzed multiple **LLM**'s for their performance on **financial arabic ner** and achieved a macro average **F1 score of 95.97**.
- Collected & standardized Arabic financial ner data and fine-tuned LLM's using **PEFT, domain, and task-specific fine-tuning** methods for NER task.

Country-level dialect identification

- Spearheaded and submitted a system for the NADI-2023 Shared task (subtask 1) in which we **fine-tuned and optimized a HuggingFace-based Arabic BERT model**.
- The **IUNADI** team implemented a customized pre-processing and ensemble method to improve model predictions for **country-level dialect identification** and reached an **F1 score of 70.22** which was the official metric of evaluation and **ranked 13th internationally**.

Identifying Sexism in Social Networks

- Contributed and led the IUEXIST team to the EXIST 2023 task 1. Our team achieved top rank among IU teams and was in the top 20 overall. The shared task focused on **identifying sexism in social networks**.
- We used **Deep Learning, Natural Language Processing, Data Preprocessing, and Ensemble Learning** techniques. We achieved an **ICM score of 71.15%** for the overall model and **ranked 9th globally**.

Driver Insurance Claims Prediction

- Utilized data mining techniques and statistical ML models to predict if a customer would file an auto insurance claim in a year.
- Multiple statistical models like **Logistic Regression, SVC, Xgboost, Adaboost, ensemble classifier using bagging** to perform the prediction and achieved a max **F1 score of 96.9** on the test data.

PUBLICATIONS

- Country-level dialect identification at NADI-2023 Shared Task
- Identifying Sexism in social networks at EXIST-2023 Shared Task