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	Sklearn, Matplotlib, Numpy, PyTorch, Tensorflow, HuggingFace, Keras, FastText, Librosa, Whisper, Nvidia-NeMO, LangChain, MLFlow, PEFT, Universal Dependencies, Prompt Engineering, GenAI
Tools	MySQL, MongoDB, Github, Tableau, Power BI, JIRA, Confluence, Linux, DevOps, CI/CD
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 - Present

- Crafted an automated **speech-to-text** script that extracts interviewee dialogue from an audio file for 13 different dialects of Spanish across 8 countries as a Data Science Fellow.
- Cross-analyzed **LLMs** to perform stt and identify the importance of vowel spacing in Spanish using **Python, Tensorflow, and PyTorch** and reduced manual annotation of data by 80%.

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

- Collaborated on developing an insurance platform called Majesco Policy Administration System for Homesite, leveraging data-driven methodologies.
- Using **SQL** developed and processed end-to-end testing for over 3 insurance products to ensure system stability and functionality.

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

- Designed a custom pandemic forecasting model stemming from **SIR** (Susceptible, Infected, Recovered) and **ARIMA** models to **predict the number of deaths due to COVID-19**. The model considered several factors including herd immunity, the number of hospital beds, and the severity of government restrictions.
- Utilized **Power BI** to summarize our findings in ten dashboards visualizing **economic impacts** on different states in the US.

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

- Designed and deployed a **True Feedback model** developed using **NLP** tools, to provide an 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.

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

- Combined **speech recognition** and **facial emotion recognition** machine learning techniques to develop an Android application to detect early onset of dementia.
- Five speech features and seven facial emotions were integrated to develop the model using **NLP, Android and Firebase**. Incorporated elements such as a facial feature extraction module, integration of speech, and facial module using NLP, Android, and Firebase.

PROJECTS

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