

Preethy P Johny

er.preethypjohny@gmail.com | +91 9811737379

<https://github.com/preethypi>

Data Scientist with an IT background, specializing in Machine Learning, Deep Learning, and Generative AI to build impactful data-driven solutions

Experience

Data Scientist, Connex AI, Manchester

June 2023 – July 2025

Language Identification using Deep Learning:

- Designed and trained a language identification model for multiple languages including minority African languages.
- Explored and gathered datasets and authenticated the ground truth, performed targeted data augmentation and detailed error analysis on the misclassified instances and lexically similar languages.
- Achieved above 95% performance across all the languages (accuracy, f1-score) on test set, including improved performance on code-switched data.
- Tools/Packages used: Tensorflow, Pytorch, Neural Network, Scikit-learn, Pandas, Numpy, matplotlib, SentencePiece, GloVe, Word2Vec, jupyter notebook, Google Cloud Translate API.
- Successfully trained and evaluated optimized models on CPU, reducing infrastructure costs while maintaining model accuracy and performance.

Topic Modelling for Thematic Content Extraction:

- Developed a probabilistic generative model to discover underlying themes in large volumes of unstructured text data.
- Conducted data preprocessing (tokenization, stopwords removal, lemmatization) to improve topic quality, used coherence score and manual interpretability evaluation metrics.
- Tools/Packages used: Gensim, SpaCy, NLTK, pyLDAvis, scikit-learn, regex, matplotlib, Word Cloud
- Enabled actionable business insights by identifying key thematic trends and aligning them with user behavior data.

Sentiment & Emotion Analysis for Multilingual Conversations:

- Developed sentiment classification models using both traditional machine learning (Naive Bayes, SVM, Bag-of-Words) and advanced transformer-based models (XLM-R, LaBSE).
- Extended sentiment analysis to detect nuanced conversational cues such as politeness and empathy.
- Conducted in-depth evaluation and bias analysis across demographic and language boundaries.
- Tools/Packages used: Tf-IDF, BoW, Scikit-learn, TensorFlow, Hugging Face Transformers, Pytorch, Pandas, NLTK, LaBSE, XLM-R, Matplotlib, Seaborn, Label Studio, LLM, Machine Learning, NLP.

Inverse Text Normalization (ITN) for Speech Processing Systems:

- Developed ITN modules to convert spoken-form text (e.g., “twenty twenty-four”) into normalized written form (“2024”) for integration into ASR systems.
- Implemented both rule-based systems using Finite State Transducers (FSTs) and neural sequence-to-sequence models for broader coverage and accuracy.
- Designed grammar rules and regex patterns for numeric, temporal, and entity normalization (e.g., dates, currencies, addresses).
- Evaluated models on multilingual utterances to ensure robustness across diverse formats and dialects.
- Improved downstream ASR output accuracy, contributing to more readable and structured transcriptions for real-time speech applications.
- Tools/Packages used: NVIDIA NeMo, PyTorch, FSTs, Python, Pandas, NumPy, Jupyter Notebook, Regex.

Predictive Analytics Dashboard with Time Series Forecasting (Master's Major Project)

- Worked towards developing an interactive web-based dashboard which can output predictive model outputs

and time-based forecasts.

- Implemented time series model using ARIMA to forecast trends.
- Tools/Packages used: Python, Flask/Django, ARIMA, Facebook Prophet, Pandas, NumPy, SQL, HTML.

EUS Engineer, Majid Al Futtaim, Dubai

Jan 2019 – Jan 2021

- Analyzed recurring IT issues and support patterns using internal ticketing systems and data-driven troubleshooting approaches.
- Created structured reports to identify trends, root causes, and predictive indicators for commonly occurring technical problems and collaborated with cross-functional teams to recommend proactive solutions based on data patterns and user behavior analytics
- Tools/Technologies: Power BI, Tableau, Excel, SQL, Python (basic), ServiceNow/JIRA

Additional Professional Experience, Various MNCs, India

Aug 2012 – Aug 2018

- **Quality Advisor, Concentrix (Feb 2016 – Aug 2018):**
Analyzed large datasets to track IT engineer performance, identify process gaps, and recommend improvements. Generated insights through data wrangling and trend analysis to support quality assurance.
- **IT Operations Associate, Accenture (Sep 2013 – Aug 2015):**
Supported Windows server migrations, patch management, and system maintenance. Analyzed system logs and performance data to troubleshoot network issues and ensure data security.
- **Technical Support Associate, IBM (Aug 2012 – Aug 2013):**
Provided frontline IT support, resolved high volumes of incidents, and maintained detailed records of issue categories for operational analysis.

Education

MSc in Data Science

Swansea University, United Kingdom

Bachelor of Computer Science Engineering

Karavali Institute of Technology, VTU University, Mangalore

Skills

Languages: Python, SQL

Skills and Technology:

Deep Learning, Machine Learning, Classification, Regression, Clustering, Statistics, NLP, Large Language Models (LLMs), Sentence Transformer, BERT, Data Science, Exploratory Data Analysis, Data Visualisation, Data Wrangling, Python and Python Packages, Django, Flask, SQL, Git, Confluence, Jira, Lotus Note, VSCode, Remedy, Excel, Power BI.

Certifications

ITIL V4 Certification

Six Sigma

AWS Certified AI Practitioner (pursuing)

References can be provided on request.