

Dhyey Patel

Austin, TX | P: +1 7373499910 | dhyeypatel1205@gmail.com

EXECUTIVE SUMMARY

Data Scientist with hands-on experience building ML systems spanning LLM safety, predictive modeling, and NLP applications. Work includes developing adversarial prompt classifiers for guardrail systems, fine-tuning transformers for detection tasks, and building NLP pipelines for large-scale semantic analysis. Proficient across classical ML, deep learning, and modern LLM workflows, always prioritizing solutions that deliver business value.

EDUCATION

UNIVERSITY OF TEXAS AT ARLINGTON

Master's of Science in Computer Science

Cumulative GPA: 3.38

Relevant Coursework: Data Analysis; Computer Vision; Machine Learning; Database Systems; Artificial Intelligence

Arlington, TX

Fall 2023 - Fall 2025

PARUL UNIVERSITY

Bachelor of Science in Computer Science

Cumulative GPA: 3.42

Vadodara, GJ

May 2018 - December 2022

WORK EXPERIENCE

BIT-INFOTECH

Data Analyst Intern

Vadodara, GJ

July 2021 – June 2022

- During my tenure as a Data Analytics Intern at BITINFOTECH, Developed and implemented end-to-end data processing pipelines, ensuring high-quality datasets for analysis through rigorous data cleaning and segmentation.
- I actively contributed to the development and validation of machine learning models to drive predictive insights for business applications.
- Synthesized complex analytical findings into actionable reports, delivering data-driven presentations to stakeholders to support strategic decision-making.

PROJECTS

JAILBREAK PROMPT DETECTION SYSTEM(LLM SECURITY)

- Developed a guardrail classifier to detect malicious inputs in Large Language Models by **fine-tuning BERT** and **DistilBERT** models on a dataset of 26,000 prompts
- Engineered a robust training pipeline on Google Colab TPUs using stratified sampling to handle class imbalance, achieving an **F1-score** of **~0.99**

FORMULA 1 RACE STRATEGY PREDICTOR

- Designed a predictive modeling pipeline to forecast Formula 1 lap times by analyzing over **100,000 telemetry data points**, including tyre degradation and weather metrics
- Achieved a peak **R² score** of **96.7%** using an optimized **Random Forest Regressor** after benchmarking performance against seven machine learning algorithms, including **LightGBM** and **Neural Networks** and Conducted comprehensive feature importance analysis to isolate critical performance drivers, providing actionable data for race strategy planning

SALON MANAGEMENT DATABASE SYSTEM

- Designed and implemented a normalized relational database schema comprising 12 tables to digitize complex salon operations, including inventory tracking and employee scheduling
- Authored advanced **SQL queries** utilizing multiple joins and aggregate functions to calculate key business metrics like **Customer Lifetime Value (CLV)** and monthly revenue
- Additionally, modeled data integrity constraints through detailed **Entity-Relationship (EER)** diagrams to ensure accurate reporting and optimized query performance

AUTOMATED SEMANTIC TOPIC CLUSTERING

- Engineered an unsupervised **NLP** pipeline to structure and analyze over 15,000 text records by leveraging Sentence-Transformers for semantic embedding and **HDBSCAN** for density-based clustering
- Integrated Generative **AI APIs (OpenAI/DeepSeek)** to programmatically generate context-aware labels and summaries for identified clusters
- Validated the approach by visualizing high-dimensional embedding spaces, successfully revealing latent topics and improving data interpretability

SENTIMENT ANALYSIS OF AIRLINE TWEETS USING BERT AND DISTIL BERT

- Engineered a comprehensive sentiment analysis pipeline to classify airline customer feedback into positive, negative, and neutral categories
- Implemented a comparative strategy by utilizing **BERT-base-uncased** for contextual embedding extraction fed into a **Deep Neural Network** classifier, versus fine-tuning a **DistilBERT** model
- Achieved **~80% accuracy** by performing rigorous text preprocessing with Regex and analyzing class distributions, demonstrating proficiency in both feature extraction and end-to-end learning

SOFTWARE/PACKAGES PROFICIENT AT

Python, Git, TensorFlow, Scikit-learn, Pandas, matplotlib, Numpy, Tableau, Excel, Machine Learning, Deep Learning, Data visualization, EDA, MySQL, Transformers, Streamlit, MLFLOW, RAG, Model Evaluation, PowerPoint Presentation, Data mining, LLMs, AI Workflow, Single & Multi AI Agents, Langchain, Pytorch, Visual Studio, Antigavity