

# Toshini Agrawal

ta2828@columbia.edu | Portfolio: learner1910.github.io | +1 6463050231 | LinkedIn: Toshini Agrawal

## EDUCATION

---

**Columbia University**, New York, USA

Aug 2025 - Dec 2026

MS in Data Science

CGPA: 3.9 / 4.0

Relevant Coursework: Finance and Structuring in Data Science, Data Structures and Algorithms, Advanced Deep Learning

**Vellore Institute of Technology (VIT)**, Bhopal, India

Jul 2019 - Jun 2023

B.Tech. in Computer Science and Engineering - Artificial Intelligence and Machine Learning

GPA: 8.77 / 10.0

Relevant Coursework: Operating Systems, Data Structures, Algorithms, AI/ML, Computer Networks, DBMS

## EXPERIENCE

---

**Columbia University**, New York, USA

Sep 2025 - Present

Graduate Research Assistant, Provenzano Lab

- Analyzed **128,000** social media posts using **Python, pandas, NLP, and sentiment analysis** to engineer language and engagement features, explaining **12-15%** of user engagement variance.
- Applied text-based feature pipelines using **computational linguistics, embeddings, and machine learning models** in **Python**, driving **8-12%** engagement improvement across **three** product lines within **six** months.
- Automated data preprocessing, visualization, and exploratory modeling workflows using **Python, pandas, NumPy and Matplotlib**, improving analysis efficiency by **30%** and reducing manual effort by **20+** minutes per experiment.

**Virtusa**, Hyderabad, India

Sep 2023 - Jul 2025

Python Developer - Full Time

- Designed a **multi-modal AI validation framework** for deepfake, copyright, and toxicity detection, cutting moderation time from **2 hours to 25 minutes** and improving efficiency by **15%**.
- Architected and deployed scalable backend systems with **Django REST, PostgreSQL, Celery, Redis, and React**, handling **40,000+** daily API requests under high concurrency.
- Developed a **RAG-based Test Case Scanner** using **GPT-4 Turbo, vector search, and retrieval pipelines**, increasing test selection accuracy by **22%** and reducing regression/API latency by **30%**.

**IIT Hyderabad**, Hyderabad, India

Feb 2023 - Jun 2023

Research Intern - Full Time

- Developed a **depression-support chat bot** with **severity-adaptive tone**, piloted with **100+** users, achieving **85% satisfaction** and measurable real-world impact.
- Implemented **NLP pipelines** including **sentiment analysis, topic modeling, and ML classification** on a **50,000+ Reddit dataset**, improving depressive cue detection accuracy by **18%**.
- Optimized **data preprocessing and feature engineering** to reduce model runtime by **35-40%**, while leading **Q and A review sessions** that increased user engagement by **30%** and pass rates by **5%**.

## PROJECTS

---

**Learning Rate Warmup Across Data Regimes**

Sep 2025 - Dec 2025

Tech Stack: Python, PyTorch, SGD, AdamW, LR Schedulers, NumPy

- Conducted **PyTorch experiments** comparing **SGD vs. AdamW** with **learning-rate warmup and schedulers**, tuning **10+ hyperparameter configurations** on **CIFAR-10, CIFAR-100, MedMNIST**, achieving **6-9% faster training**.
- Leveraged custom **NumPy training loops** to assess optimizer behavior and analyze **few-shot and full-data regimes** across three dataset scales, reducing training instability by **20%**.

**PlagBench: LLMs for Plagiarism Detection - ARR/EMNLP**

Nov 2023 - Jun 2024

Tech Stack: Python, Scikit-learn, NLTK, SpaCy, TensorFlow, Keras, MTurk.

- Spearheaded evaluation of an **LLM benchmarking system** on paraphrased text, testing **5+ transformation types**, improving evaluation reliability by **25%** compared to human-labeled data.
- Validated **Python NLP pipelines** using **SpaCy and TensorFlow**, increasing model accuracy by **5%** and evaluation reliability by **30%**.

**Traffic Crash Severity - Springer LNCE**

Feb 2023 - Aug 2023

Tech Stack: Python, scikit-learn, XGBoost, Data Analysis, Machine Learning.

- Modeled **injury and property-damage severity** using real-world traffic and environmental data, training and evaluating **4+ machine learning models** including regression and ensemble methods.
- Conducted **feature engineering and model evaluation** during a research internship at **IIT Jodhpur**, identifying **10+ high-impact risk factors** and improving prediction performance by **15%**.

## TECHNICAL SKILLS

---

**Languages:** Python, C++, JavaScript, SQL, Bash, Java

**Frameworks:** Scikit-learn, NLTK, SpaCy, TensorFlow, Keras, Django, OpenAI GPT, MERN, PyTorch

**Tools:** Linux, UNIX, Docker, Git, PostgreSQL, MySQL, SQLite, AWS, Agile

**Data Science Concepts:** Machine Learning, Deep Learning, Natural Language Processing, Big Data Technologies, Predictive Analytics, Prompt Engineering, Artificial Intelligence, Business Analytics, Database Management