Harish Upadhyay
Email: Harish.Upadhyay@gmail.com | Location: Remote | Experience: 3 years

SUMMARY

Machine Learning Engineer (NLP & Retrieval) with 3 years of hands-on experience building production semantic search, retrieval, and RAG systems. Strong Python, transformer-based NLP, embeddings, and deployment experience (Streamlit, FastAPI). Proven track record integrating LLMs, sentence-transformers, FAISS/Chroma, and cloud deployment pipelines.

TECHNICAL SKILLS

- Languages: Python (daily), SQL, Bash
- NLP / ML: Transformers (Hugging Face), sentence-transformers (all-MiniLM-L6-v2), tokenization, sequence labeling, classification, entity extraction, sentiment analysis
- Retrieval & Vector DBs: FAISS, Chroma, Qdrant, vector embeddings, ANN search, dense–sparse hybrid retrieval
- LLMs & RAG: OpenAI / OpenRouter usage patterns, prompt engineering, retrieval-augmented generation (RAG), LLM orchestration
- Frameworks & Tools: PyTorch, scikit-learn, LangChain (basic), Streamlit, FastAPI, Docker, Git, CI/CD (GitHub Actions), AWS (S3, EC2, ECR)
- Evaluation & Explainability: ranking metrics (MRR, NDCG), precision/recall/F1, embedding tuning, feature importance and explainability notes

EXPERIENCE

Machine Learning Engineer — NovaTech (Contract) — Remote — Jan 2023 – Present

- Designed and implemented a semantic search pipeline using sentence-transformers (all-MiniLM-L6-v2) to produce high-quality embeddings and FAISS-backed nearest-neighbor retrieval for enterprise search.
- Built a RAG architecture combining vector retrieval (Chroma/FAISS) + LLM prompt templates to answer business queries with provenance and source attribution.
- Developed FastAPI microservices to serve embeddings and retrieval endpoints; created Streamlit demo apps showcasing ranking and evaluation dashboards.
- Tuned embedding dimensionality and experimented with hybrid dense–sparse retrieval to improve MRR and NDCG by 18% on internal datasets.
- Implemented CI/CD with GitHub Actions, Dockerized services, and deployed models and search infra to AWS (ECR + ECS / EC2).
- Collaborated with product teams to define KPIs, monitoring, and model performance evaluation; wrote modular, well-tested Python code and version-controlled pipelines.

NLP / ML Engineer — DataWorks Lab — Jul 2021 – Dec 2022

- Built text classification, NER, and sentiment analysis models using transformer fine-tuning and classical baselines.
- Integrated sentence-transformers for embedding generation and created pipelines to index embeddings in Chroma for semantic matching.
- Created evaluation suites measuring precision/recall, F1 and ranking metrics; iterated on preprocessing and embedding normalization strategies.

SELECTED PROJECTS

- Enterprise Candidate-Matching System — End-to-end: preprocessing, embedding generation with sentence-transformers, FAISS/Chroma indexing, retrieval API (FastAPI), RAG-based explanation layer using an LLM. Deployed on AWS.
- Streamlit Search Console — Interactive UI showing query embeddings, nearest neighbors, similarity scores, NDCG/MRR charts and fine-grained document provenance.
- Embedding Tuning Experiments — Compared MiniLM and other models, tuned pooling and normalization to improve retrieval accuracy. Documented experiments and hyperparameters.

EDUCATION

PGDM — Analytics Specialization — Birla Institute

OPEN SOURCE / OTHER

- GitHub: github.com/savyapandey-ml (examples: embedding pipelines, Streamlit demo, FastAPI retrieval service)
- Regularly use tools: Git, Docker, AWS, experiment tracking, test-driven development