Binay Kumar Pradhan

AI Engineer | Founding Engineer | Software Engineer

PERSÖNLICHE DATEN

Adresse	Cottbus, Deutschland
E-Mail	pradhan.binay.de@gmail.com
LinkedIn	linkedin.com/in/i-binay
GitHub	github.com/vinayin
Website	vinayin.github.io
Verfügbarkeit	Sofort

BERUFLICHES PROFIL

Masterstudent der Künstlichen Intelligenz mit Schwerpunkt auf Modellkompression (Pruning, Quantisierung, Knowledge Distillation), Reinforcement Learning, ML-Lösungsdesign und Large Language Models (LLMs). Mehr als vier Jahre Berufserfahrung in MLOps und Software Engineering, einschließlich Backend-Entwicklung, Data Science sowie der Entwicklung und Bereitstellung skalierbarer End-to-End-Machine-Learning-Pipelines auf den Cloud-Plattformen AWS und GCP.

Ziel ist eine Position als AI Engineer, in der produktionsreife Lösungen mit verbesserter Leistung, Kosteneffizienz und Robustheit entwickelt werden. Dabei soll ein Beitrag zu nachhaltiger Entwicklung durch ethische KI-Anwendungen geleistet werden.

BERUFSERFAHRUNG

Wissenschaftlicher Mitarbeiter
06.2023 – 03.2025

Brandenburgische Technische Universität (BTU) Cottbus-Senftenberg | Cottbus, Deutschland

- Entwicklung eines sprachübergreifenden Datenvisualisierungstools mit PyQt6 durch Integration von Python-, Java- und MATLAB-Subroutinen sowie Automatisierung von MongoDB-Datenextraktionspipelines, was Forschungsprozesse erheblich beschleunigte.
- Konzeption und Umsetzung verteilter Reinforcement-Learning-Simulationspipelines zur Optimierung von Supply-Chain-Prozessen (insbesondere Lagerhof-Management) durch Formulierung als Zeit-Kosten-Scheduling-Problem und Einsatz von Constraint Programming. Ableitung effizienter und nachhaltiger Logistikstrategien, skalierbar für großflächige Distributionsszenarien.
- Steigerung des Teilnehmer-Durchsatzes um 50 % durch Entwicklung gezielter EEG-Kalibrierungsparadigmen und Automatisierung von Datenanalyse-Pipelines, wodurch manuelle Engpässe in datenintensiven Forschungsumgebungen beseitigt wurden.
- Entwicklung und Veröffentlichung des *pyETA-toolbox* auf PyPI, eines produktionsreifen Open-Source-Toolkits für Echtzeit-Fixationserkennung und Eye-Tracker-Validierung in der passiven BCI-Forschung (pypi.org/project/pyETA-toolbox).

AI Data Engineer II
06.2022 – 09.2022

Wolkus Technology Solutions Pvt Ltd | Bangalore, Indien

- Konzeption und Deployment von ETL-Pipelines auf AWS SageMaker mit vollständiger Migration eines GCP-Datenmarts nach AWS sowie Verbesserung der End-to-End-Beobachtbarkeit durch Elasticsearch-Integration zur Unterstützung skalierbarer ML-Trainingsworkflows.
- Reduzierung der p95-API-Latenz und des AWS-Lambda-Cold-Start-Overheads um 20 % durch Refactoring des serverless Codebase und Optimierung der Cloud-Compute-Effizienz für KI-gestützte Produktionsworkloads.
- Entwicklung eines Loggers und Veröffentlichung der Open-Source-Bibliothek *pylogger-slack*

auf PyPI für strukturierte Logging mit Echtzeit-Slack-Integration in produktiven ML-Pipelines (pypi.org/project/pylogger-slack).

Software Engineer II 10.2021 – 03.2022
Myelin Foundry Pvt Ltd | Bangalore, Indien

- Entwicklung von AI-Funktionen für In-Car-Dashboards zur Gesichtsemotionserkennung, Handgestenerkennung und Objekterkennung mit dem MediaPipe-Computer-Vision-Framework.
- Deployment produktionsreifer Deep-Learning-Modelle auf automotiven Edge-Geräten durch Anwendung von Modellkompressionstechniken (Quantisierung und Pruning), wodurch niedrige Latenzzeiten auf ARM-Architektur erreicht und fundierte Expertise in kosteneffizienten Implementierungen aufgebaut wurden.

Data Science Engineer 07.2019 – 10.2021
Utopia India Pvt Ltd | Bangalore, Indien

- Steigerung des Kundenengagements um 30 % im Zeitraum Q1–Q3 2021 durch Entwicklung von REST-API-Modulen, Beseitigung von Architektur-Engpässen und Verbesserung der Datenvisualisierungen im Rahmen des 4C-Projekts (Classify, Collect/Scrape, Correlate/Map, Clean/Standardize).
- Entwicklung eines Proof-of-Concept für ein Reinforcement-Learning-Modell zur automatisierten Text-Labeling mit Echtzeit-Human-Feedback.
- Erstellung eines weiteren Proof-of-Concept für industrielle Textgenerierung mit BERT und GPT zur Anwendung auf freiformatierte Texte.
- Erreichung einer Extraktionsgenauigkeit von 85 % und einer Effizienzsteigerung um das 15-fache gegenüber manuellen Prozessen durch Entwicklung eines regex-basierten Information-Extraction-Parsers mit Konsensalgorithmen.
- Wartung von CI/CD-Pipelines, strukturierter Logging-Observability sowie kostengünstiger Verteilung von AWS-SPOT-Instanzen für nahtlose verteilte ML-Workflows.
- Entwicklung eines Bildabgleich-Algorithmus zur automatisierten PDF-Textsegmentierung.
- Leitung der Anforderungsaufnahme und cross-funktionalen Koordination für ein 10-köpfiges ML-Team im 4C-Datenbereinigungs- und Standardisierungsprojekt.

Machine Learning Intern 01.2019 – 07.2019
Utopia India Pvt Ltd | Bangalore, Indien

- Konzeption und Evaluierung von Text-Matching-Algorithmen mit vortrainierten Word-Vektoren, eigenen Embeddings, Fuzzy Logic und Mengenrelationen zur Verbesserung der Entity-Matching-Genauigkeit.
- Entwicklung datengetriebener NLP-Algorithmen zur automatisierten Standardisierung und Strukturierung freiformatierter Text-Labels und -Werte.
- Aufbau einer Batch-Datenbereinigungspipeline mit Makefile-gesteuerten Workflows für skalierbare Großdateiverarbeitung.

Research Intern 05.2018 – 07.2018
Institute of Development and Research in Banking Technology (IDRBT) | Hyderabad, Indien

- Erstellung und Annotation hindi-bengalischer Code-Mixed-Datensätze über die Twitter-API, was zu einer Veröffentlichung in ACM TALLIP (2020) mit F1-Score-Verbesserungen von 5–6 % beitrug und grundlegende Expertise in linguistischer Datenverarbeitung aufbaute.

PROJEKTE

TechEU Editor 07.2025
Tech Europe Hackathon | Berlin, Deutschland

- Konzeption einer Multi-Agent-Orchestrierungspipeline mit CrewAI und OpenAI durch Einsatz spezialisierter Agenten zum autonomen Scrapen von Quellen, Generieren strukturierter Artikel, Verwalten von Zitationen und gezielter Textanpassung für eine vollständige redaktionelle Automatisierung (github.com/VinayIN/tech-europe-hackathon).

HR Assist System 03.2025 – 07.2025
Erasmus+ Entrepreneurship Labs 2025, Universität Catania | Catania, Italien

- Entwicklung und Prototyping eines KI-gestützten B2B-SaaS-Produkts durch Anwendung unüberwachter Clustering-Algorithmen auf Mitarbeiter-Performance-Reviews, Skill-Profile und Verfügbarkeiten zur automatischen Erstellung optimal ausbalancierter Teamkonfigurationen.
- Präsentation des Produktkonzepts im Rahmen des Erasmus+-Programms vor Investoren und Mentoren mit dem Ziel der Umsetzung als studentisches Start-up.

AUSBILDUNG

M. Sc. Künstliche Intelligenz 10.2022 – voraussichtlich 2026
Brandenburgische Technische Universität (BTU) Cottbus-Senftenberg | Cottbus, Deutschland

- Note: 2,6
- Masterarbeit (in Bearbeitung)

B. Tech. Information Technology 08.2015 – 06.2019
International Institute of Information Technology Bhubaneswar (IIIT-B) | Bhubaneswar, Indien

- Note: 2,8
- Thesis: *Mapping unstrukturierter Texte in strukturierte Texte* – NLP-basierte Verfahren zur Transformation freiformatierter Texte und Strukturierung anhand einer Vorlage.

FACHKENNTNISSE

Programmiersprachen: Python, Java, MATLAB, Bash
ML/AI-Frameworks: PyTorch, TensorFlow, Hugging Face, Scikit-learn, ONNX, MediaPipe
Generative KI & LLMs: CrewAI, Transformers, Retrieval Augmented Generation (RAG), Agentic AI
AI/ML-Spezialisierungen: Deep Learning, Natural Language Processing (NLP), Computer Vision, Modellkompression (Pruning, Quantisierung, Knowledge Distillation), Reinforcement Learning, Spracherkennung, Brain-Computer Interface (BCI)
MLOps & Cloud: MLflow, Docker, CI/CD, Git, REST APIs, Makefile, AWS (SageMaker, Lambda, SPOT Instances), GCP (Firebase)
Datenbanken: MongoDB, Elasticsearch, PostgreSQL, Weaviate, Supabase
Datenvisualisierung: Streamlit, Gradio, PyQt6

SPRACHEN

Englisch	C2 — Fließend
Deutsch	A2 — Grundkenntnisse

VERÖFFENTLICHUNGEN

Bhattu, S. N., Nunna, S. K., Somayajulu, D. V. L. N., und **Pradhan, B.** (2020). Improving Code-mixed POS Tagging Using Code-mixed Embeddings *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, Vol. 19, No. 4, Article 50, S. 1–31. Veröffentlicht: 29. März 2020
DOI: 10.1145/3380967
Vorschlag eines hierarchischen tiefen rekurrenten neuronalen Netzes mit CRF-Layer zur Verbesserung des POS-Taggings von Code-Mixed-Social-Media-Texten (Bengalisch-Englisch, Hindi-Englisch, Telugu-Englisch). Entwicklung eigener Code-Mixed-Word-Embeddings; Erzielung von F1-Score-Verbesserungen von 5,81 %, 5,69 % und 6,3 % gegenüber CRF-Baselines bei den drei Sprachpaaren.