Chetan Khadke

New Jersey, USA

khadkechetan@gmail.com

SKILL SETS

Area:
Natural Language processing
Large Language Model
Text Mining/Classification
Information Retrieval
Machine Learning
Deep Learning

Language:
Python, R, Java

PROFESSIONAL PROFILE

- Data Scientist with over 10 years of experience, specializing in Natural Language Processing (NLP) and Machine Learning (ML) for solving diverse and complex challenges across industries.
- Recent expertise includes advanced information extraction, text classification, and clustering using state-of-the-art Large Language Models (LLMs) such as Transformers, Llama-Index, LangChain, and Gemini 1.5 Pro.
- Independently developed and deployed multiple Proof-of-Concepts (POCs) showcasing scalable NLP solutions, including integration with cloud platforms (AWS, GCP) and containerization for seamless deployment.
- Proficient in Python, R, and Java, with recent experience in using modern ML libraries like Hugging Face Transformers, LLM, and vision LLM to tackle NLP tasks.
- Published and presented four research papers in prominent international conferences, sharing novel insights and methodologies in NLP and ML applications.
- Actively contribute to the Data Science community by mentoring, sharing research, and participating in forums, ensuring continuous learning and staying updated on the latest AI trends.

Packages/Tools: Transformers, Llama-Index, LangChain, Gemini 1.5 pro, Llama-cpp, Spacy, Keras, LayoutLM, Qwen LM, Donut, UDOP, OpenAI, MLflow, Bert, PyTorch, fasttext, NLTK, sentence-transformer, bertopic, top2vec

MORE SKILL SETS

Database: MySQL, MongoDB, Neo4j, Elastic Search
Machine Learning: Logistic Regression, Maximum Entropy Model, Linear Regression, SVM, Neural Network, boosting (Ada Boost, Gradient boosting Machine, Extreme Gradient Boosting (XGB), Decision Tree, Random Forest, PCA, LDA, Gaussian Mixture model, Graphical model of ML, K-Means, Hierarchical Clustering, Density Based Clustering.
NLP: Large Language model (LLM)(Llama 3.2, Gemini 1.5 pro, Mistral AI), Visual LLM(Llava, Qwen 2), LayoutLM, Donut, UDOP, Text Classification and Clustering, Named Entity Recognition(NER), Entity Linking, Topic Modeling, Named Entity Linking
Deep Learning: Convolution Neural Network, Recurrent Neural Network, Long Short-Term Memory Neural Network, Encoder-Decoder Neural Network, Transformer
AWS: ECR, ECS, S3, CloudWatch, SageMaker, Inference Endpoints
ML Deployment: Docker container, Managing docker container with ECR, Git, ML flow

PROFESSIONAL EXPERIENCE

Xoriant Solutions
(Aug 2024 - Current)

Information Extraction and Réconciliation – Xoriant, USA

The solution is dedicated to extracting critical information with utmost accuracy from trade documents sourced from various channels, including contracts, invoices, bills of lading, and financial statements. Designed to ensure data consistency, this solution performs robust reconciliation across multiple sources, identifying and reporting discrepancies in real time to mitigate potential financial losses. The solution integrates seamlessly into financial

workflows, allowing stakeholders to address discrepancies proactively and make data-driven decisions with confidence.

Responsibility:

- Created and curated prompts leveraging the latest Gemini Pro 1.5 for optimized information extraction, ensuring high precision and relevance.
- Developed strategies to mitigate hallucination risks by crafting intelligent, context-aware prompts for reliable output.
- Designed advanced reconciliation status checks using Large Language Models (LLMs) and sentence-matching algorithms to improve data accuracy and consistency.
- Conducted defect validation and implemented bug fixes to enhance system stability and performance.

Xoriant Solutions
(May 2021 -July 2024)

Information Extraction from Documents (Document AI) – Xoriant, Pune

Document AI is a robust, scalable solution for automatic document type detection and field/entity extraction. Capable of classifying over 100 document types and extracting more than 50 unique fields, Document AI allows the addition of custom templates to accommodate diverse formats. It efficiently processes millions of documents weekly, supporting not only text fields but also checkbox and radio button detection, address parsing, and signature recognition. Integrated with a rules engine, Document AI streamlines business processes by automating comprehensive document handling tasks.

Responsibility:

- Spearheaded the development of a Named Entity Recognition (NER) module, leveraging a variety of NER libraries to optimize performance and accuracy.
- Actively engaged in exploring and setting up client-centric annotation tools, tailoring solutions to meet specific client requirements.
- Led the design and training of customized models based on layoutLM, Donut and UDOP frameworks, enhancing their efficiency and applicability
- Successfully implemented a Donut-based DocVQA and LLM based approach, demonstrating proficiency in advanced document understanding techniques
- Achieved a remarkable 70% improvement in data extraction accuracy 3 through the implementation of innovative, custom-tailored solutions.
- Managed both the coding and deployment pipelines, ensuring seamless integration and functionality of machine learning models
- Expertly deployed machine learning models in containerized environments, employing best practices for scalability and maintenance

Persistent Systems
(Apr 2019- May 2021)

Knowledge Graph – PSL, Pune

Created comprehensive knowledge graphs to establish connections between organizations, individuals, and content from diverse data sources, enabling enhanced data insights and relationship discovery

Responsibility:

- Conducted experiments with various Named Entity Recognition (NER) libraries to optimize entity extraction performance.
- Designed and implemented classification models for advanced relation extraction tasks, enhancing NLP model capabilities.
- Developed custom NER solutions tailored to project-specific requirements, improving accuracy and relevance in entity recognition.
- Executed semantic matching algorithms to enhance contextual understanding and similarity assessment within data.
- Performed Name Entity Linking (NEL) using SpaCy and Wikipedia resources to create robust entity relationships and improve model precision.
- Managed and optimized database operations with Neo4j, ensuring efficient data retrieval and storage for NLP applications.

- Deployed machine learning models in containerized environments to streamline scalability and integration within production workflows.

Intent Predication – PSL, Pune

(Aug 2018- Feb 2019)

Focused on enhancing customer satisfaction by developing an automated chatbot capable of delivering accurate, relevant, and timely responses to user queries.

Responsibility:

- Analyzed and interpreted complex data, sharing key insights with the team to inform strategic decision-making.
- Implemented various classification algorithms, including SVM, Naive Bayes, and RNNs, as well as distance-based methods, to accurately classify intents.
- Applied sentence-based similarity metrics to effectively capture semantic relationships, enhancing the model's ability to understand nuanced meanings.

Extraction of structured and meaningful information from ticket dumps, TCS, Pune

Tata Consultancy Services

(2016- July- 2017)

Focused on extracting relevant information from extensive ticket data sets, followed by time series analysis to compute key performance metrics and trends, as part of Ignio™ team.

Responsibility:

- Applied a range of NLP techniques to efficiently process and analyze unstructured data, extracting valuable insights.
- Conducted experiments with various clustering and classification techniques to group similar events effectively.
- Contributed to the development of advanced text-matching algorithms, enhancing the accuracy of similarity detection in text data.

Mining workflow from given System Integration Test cases, TCS, Pune (2014-2016)

(2014-2016)

Developed a mining system model from a set of test cases, aiming to derive a comprehensive understanding of the software system based on real-world test scenarios

Responsibility:

- Utilized various NLP techniques for effective preprocessing of input data, ensuring high-quality data preparation for analysis.
- Applied machine learning algorithms to classify test cases into predefined categories, improving data organization and accessibility.
- Developed a standalone application using Java Swing to demonstrate project outcomes to clients, enhancing user experience and presentation.
- Published four research papers in prestigious international conferences, contributing to advancements in the field and gaining industry recognition.

TCS-Initial Learning Program (ILP), TCS, Pune

(Nov 2013-Jan-2014)

TCS-ILP emphasizes comprehensive training for new joiners, covering technical skills, soft skills, client communication, and professional etiquette.

EDUCATION

(2009 – 2013)

Dharmsinh Desai University, Nadiad, Gujarat

Bachelor of Technology in computer science with GPA: 8.5/10, 2009-2013

ACHIEVEMENTS

- I have presented my work at the 11th International Conference on Evaluation of Novel Approaches to Software Engineering on 27-28 April 2016, Rome Italy.
- Bravo Award was received 3 times for Excellent work in Persistent System.
- XOR-Champ Award was received 1 time for Excellent work in Xoriant Solutions.

CERTIFICATIONS

- "Machine Learning" – Secured silver medal in course from Indian Institute of Technology – Madras (IIT-M) offered from NPTEL.
- "Deep Learning" – Indian Institute of Technology – Madras (IIT-M) offered from NPTEL.

RESEARCH PAPER DETAILS [1]

2015

1. Chetan Khadke, Sunjit Rana and Vipul Shah "Recovering Workflows from Functional Tests" in 23rd IEEE International Conference on Program Comprehension at Florence Italy.
2. Vipul Shah, Chetan Khadke and Sunjit Rana "Mining process models and architectural components from test cases" Software Testing, Verification and Validation Workshops (ICSTW), 2015 IEEE Eighth International Conference on 13-17 April 2015 at graz.
3. Vipul Shah, Chetan Khadke and Sunjit Rana "Knowledge Transition: Discovering models from functional tests" in the 26th IEEE International Symposium on Software Reliability Engineering on 2-5 November 2015 at GAITHERSBURG, MD, USA.

2016

4. Presented tool demonstration "Knowledge Acquisition (KnowX)", tool demonstration in the 11th International Conference on Evaluation of Novel Approaches to Software Engineering on 27-28 April 2016, Rome Italy.

BLOG/PAPER LINK

1. Paper link: http://dblp.uni-trier.de/pers/hd/k/Khadke:Chetan
2. Medium blogs: https://github.com/khadkechetan/khadkechetan?tab=readme-ov-file#open_book-articles