NIKHIL SHARMA

B.Tech, University of Delhi

Mobile: +91-9555369618
Email:nikhilsharma129540@gmail.com

EDUCATION

Bachelor of Technology
Electronics Engineering
University of Delhi

Master of science
ML&AI
Liverpool john moores university

SKILLS

- Programming Languages: Python, Shell
- Databases: SQL
- Cloud computing: Azure
- Tools: Databricks, Elasticsearch, Redis, Docker, Streamlit

CERTIFICATES/ARTICLES

- Natural language processing with probabilistic models, *deeplearning.ai*
- Natural language processing with classification and vector spaces, *deeplearning.ai*
- NLP based information retrieval system, *Medium (Towards data science)*

PROFESSIONAL EXPERIENCE

Associate Scientist
AXA XL
July 2021 – Present

Integrated Solution for Property & Casualty (P&C) Insurance Risk Management [POC+MVP]

- Developed microservices based integrated solution aimed at optimizing risk management through address extraction, validation and parsing and natural hazard identification
- Utilized polynomial boundaries of countries to randomly select latitude and longitude coordinates, employing reverse geocoding to obtain a diverse set of addresses for model finetuning to extract addresses
- Fine-tuned transformer-based language models for address extraction, including DistilBERT, BERT, and RoBERTa, xlm-RoBERTa, attaining f1 score ~95%.
- Leveraged Melissa address validation APIs to convert the extracted addresses into precise latitude and longitude coordinates, which were then integrated with Geographic Information Systems (GIS) to obtain detailed natural hazard information for each address.

Cyber Submission Triage for Efficient Submission Assessment [POC+MVP]

- Developed an end-to-end Python framework that seamlessly integrates the entire process for submission triage, including preprocessing, data extraction, postprocessing, validation, rule-based mapping, and output report generation.
- Created document classification methods by fine-tuning transformer-based models like layoutLM, as well as logo detection and keyword searches for document identification.
- Developed proprietary methods for extracting tables from PDF documents using layout information using Camelot, tabula, and pdfminer.
- Integrated a comprehensive set of validations and rules for extracted data quality assessment and categorization.
- Integrated scoring system mapping responses to multiple severity levels such as Poor, Below Average, Average, Above Average, and Strong, facilitating an insightful and structured evaluation of submissions.

HR Document Chatbot for Real-Time Query Resolution using openai LLMs [POC]

- Build an AI-driven chatbot using openai LLMs, vector databases, custom prompts, and ReAct agents to understand and resolve the query
- Integrated memory and history tracking mechanisms to enable the chatbot to have context-aware conversations and deliver more relevant responses.
- Ensured the chatbot maintains a friendly tone in conversations while strictly adhering to providing responses derived from HR policy documents.

Clause Search and Discrepancy Identification in Policy Documents [POC]

- Developed an advanced system for clause search and discrepancy identification in policy documents by leveraging paragraph similarity algorithms.
- Utilized Universal Sentence Encoder (USE) and BERT embeddings to represent document chunks, enabling high semantic understanding and accuracy.
- Incorporated cosine similarity measures to identify and match similar clauses, enhancing precision in clause search.
- Applied the diff-match-patch algorithm for detecting discrepancies between clauses in policy documents.

	• Integrated fuzzy matching techniques to capture near matches and variations in text, further refining the discrepancy identification process. **Document Search Engine Using Azure Cognitive Search API [POC]** • Build a robust document search engine utilizing Azure Cognitive Search API for intelligent search capabilities on insurance documents. • Implemented document indexing through a microservices-based architecture using FastAPI. • Build backend APIs for user-friendly UI to facilitate seamless interaction with the search engine, improving user experience and accessibility. **OCR Engine for PDF Documents Using Microservices-Based Architecture [POC+MVP]** • Build OCR engine capable of processing both searchable and non-searchable PDF documents using Google Tesseract OCR. • Employed a microservices-based architecture using FastAPI, ensuring a scalable and modular system. • Integrated Celery for parallel processing, significantly enhancing the engine's performance and efficiency in handling large volumes of data. • Containerized the application using Docker for ease of deployment, versioning, and environment consistency. • Deployed the solution in a production environment using OpenShift, ensuring high availability and seamless scaling as per demand.
Lead Associate **WNS (GSK)** July 2020 – August 2021	**Text Analytics Tool for Insight Extraction at GlaxoSmithKline [POC+MVP]** ● Developed an industry-scalable text analytics solution by designing the architectural framework and leading requirement gathering ● Developed a rule-based methodology to identify topics/themes in medication-related talks between Medical Science Liaisons (MSLs) and Healthcare Professionals (HCPs). ● The output from the rule-based model was used to train and fine-tune a BERT model, resulting in a classification accuracy of 79%. ● Built an end-to-end pipeline that included data extraction, processing with Databricks, and visualization with Power BI, reducing human intervention to 10%.
Business Analyst **EXL (Liberty mutual, UNUM)** Nov 2018-July 2020	**Call Load Analysis at Liberty Mutual's Contact Centre [POC]** ● Developed an ETL pipeline to take audio recordings from Amazon S3 and convert them to text using AWS transcription services. ● In accordance with GDPR, algorithms were developed to hide personally identifying information. ● For data preprocessing and purification, I used NLP packages like Spacy and NLTK, as well as regular expressions. ● NLP techniques such as n-grams, word frequency, and word cloud were used for exploratory data analysis. ● Gensim was used to identify prominent topics in the data using LDA and NMF topic modelling techniques. ● Developed regex-based models for label generation to apply supervised techniques. ● Using PyTorch, I fine-tuned a BERT model for multi-label classification, attaining an accuracy of roughly 84% **Claim Prediction Model for UNUM [POC]** ● Executed the POC of the claim prediction model that led to a full-time project. ● Build a logistic model to help benefits team in claim duration predictions for minimizing overpayments. Text mining has been carried out on unstructured claim notes to extract new variables that can be clubbed with structured data for modeling purposes. ● Savings were reported in the range of 300k-700k USD based on varying threshold levels
Programmer **Junati innovations** Jun 2017- Nov 2018 *Machine learning, signal processing*	**Prediction of Arrhythmias Using ECG Analysis [Research]** ● To minimize noise in ECG readings, digital signal processing techniques such as the Fast Fourier Transform were used. ● Principal Component Analysis was used to improve data quality and ensure the retention of vital information. ● In-depth research on heart rate variability analysis for stress evaluation was conducted. ● Developed a neural network model for accurate classification of different arrhythmia types.