# Shivam Singh

**Data Scientist**

Pragmatic, competitive, and efficient Data Scientist with more than 2.5 years of experience in developing, validating, and deploying Machine Learning and Deep Learning models. My primary focus was on Natural Language Processing using state-of-the-art transformer models, and my domain of work was in healthcare and legal.

## Work Experience

**Hobasa India Pvt. Ltd, Pune, India –** ***March 2022 – present***
**Data Scientist**

**Rule Extraction and Execution for Rule Engine:**

- Developed and implemented project focused on rule extraction from unstructured text data, utilizing natural language processing (NLP) techniques and an entity extraction system to identify key entities and relationships within the text.
- Incorporated feedback from domain experts and iteratively improved the rule extraction and entity extraction models, resulting in enhanced accuracy and efficiency.
- Built a pipeline to prepare the data for training the Flair and Blackstone model using prodigy and fine-tuned Blackstone and Flair model for entity recognition on ECFR data for custom entities, that was not present in the flair model.
- After identifying entities and verifying conditions and actions present in a sentence, we utilize a Language Model (LLM) to generate attributes and Python code based on the text and features present in client data.
- Mapping of data with attributes has been completed, and Python code has been executed based on a triggered schedule. The logged errors are stored in Mongo-DB for tracking purposes, and execution plots are generated to display the process.

**Chat-Bot end to end:**

- Collaborated closely with Subject Matter Experts (SMEs) to define intents, entities, and custom context for a domain-specific Chat-Bot.
- Developed three versions of the Chat-Bot, utilizing different architectures, including the default DIET architecture, Distil-BERT, and generative AI models like GPT-2 and GPT-3.
- Designed and implemented the Chat-Bot using the Rasa framework, enabling it to identify and respond to custom context within the domain-specific text corpus.
- Implemented deployment of the Chat-Bot using NGINX as a reverse proxy server to efficiently manage incoming requests and distribute them to gunicorn, ensuring optimal performance and scalability.
- Configured gunicorn to serve the Chat-Bot application, utilizing multiple worker processes to handle concurrent requests effectively and maintain responsiveness during peak loads.

**Email Classification & Integration:**

- Developed an industry-specific text classification model for email categorization, enhancing accuracy in categorization and operational efficiency.
- Deployed the classification model on both local instances and Outlook web instances, enhancing email management capabilities.
- Extended the pipeline to automatically identify emails with attachments and seamlessly connected them to a document extraction pipeline, consolidating information from attachments into a unified platform.
- To perform extraction from attachments using LLM model, considering various types of attachments such as PDF, PNG, JPG, Excel, Word, Text, etc.
- Achieved optimal performance by optimizing and fine-tuning Large Language Models (LLMs), such as BERT, for email classification, resulting in more than 90% accuracy.
- After each process on web Outlook, it automatically creates a dedicated folder and moves all the mails related to classification into their respective folders.

## CONTACT


- +91-7007320632
- shivamcse17818@gmail.com
- https://www.linkedin.com/in/shivam-s-634759169/


## SKILLS

**Techniques:**

- Document Question Answering
- LLM fine-tuning
- Entity Extraction from text
- Intelligent Document Processing (IDP)
- Text Summarization - Transformers
- Similarity Search –Transformers
- Fine tuning NLP Transformers using In-house data
- Exploratory Data Analysis (EDA)
- Clustering and Topic modeling

**Data Annotation Tools:**

- Prodigy – for text annotations
- UBIAI – for document annotations

**Tools, Frameworks and Libraries:**

- Python
- scikit-learn
- PyTorch
- Data Analytics
  - Pandas
  - NumPy
  - Seaborn
  - Matplotlib
- Databases
  - MySQL
  - PostgreSQL
  - Mongo-DB
- NLP
  - NLTK
  - Spacy
  - transformers
  - PyTorch
  - Tensor-Flow
  - Keras

- Flask & Django for creating API's
- Docker – for productionizing the API's
- Rasa – Chabot
- Azure Databricks
- Hugging face for transformer models

**Auto Analytics & their deployment:**

- Created an auto-analytics pipeline that can be integrated with Mongo-DB. It has the capability to directly ingest both CSV and Excel files.
- In the background, it utilizes Open-AI's LLM models to process the data, extract insights, and generate goals for plotting insights derived from the data.
- First, we attempted to generate code and execute it for Seaborn and Matplotlib, provided by the Lida library. However, for the final deployment, we opted to write our own custom code generation for Plotly and executed it on the data.
- Used Stream-lit to create a UI, and deployed it on a development server using a docker container, accessible through a URL for SME's to use.

## Work Experience

**3 Analytics Pvt. Ltd, Chennai, India** **– Jun 2021 – Dec 2021**

**Data Scientist**

**Social Media Monitoring:**

- **Business Objective:** Monitored social media, particularly Twitter, to detect and categorize adverse events resulting from COVID vaccination and other drugs, while identifying genuine tweets based on predefined criteria.
- **Data Processing:** Extracted and preprocessed Twitter data, applying text cleaning techniques to ensure high-quality analysis.
- **NLP Analysis:** Employed advanced natural language processing techniques, such as cluster analysis and topic modeling, to uncover patterns in social media conversations. Compared the performance of multiple transformer models, including Bio-BERT and Clinical-BERT, and devised a strategy for identifying symptoms in the text.
- **Classification:** Leveraged labeled data to train a Bio-BERT model for tweet classification into adverse event and non-adverse event categories, achieving an impressive F1-score of 66.6%.
- **Deployment:** Connected with the Twitter API to gather real-time data and began performing classification on live tweet data, storing the results into a MySQL database to provide insights post-classification.

**Email Monitoring:**

- **Business Objective:** Classified incoming emails, including attachments, into multi-class categories such as Adverse Events, Product Quality, Complaint Medical Inquiry, and Others (Non-actionable, Spam, or Junk emails).
- **Approach:** Utilized the SMTP library to connect to the company's Outlook server, enabling real-time email retrieval. Annotated email data in collaboration with Subject Matter Experts (SMEs) for accurate classification.
- **Model Development:** Trained specialized transformer models, including Bio-BERT and Clinical-BERT, using ML flow for deployment, achieving high accuracy in classifying emails.

## Achievements:

- My **Kaggle** journey has been marked by active participation in a diverse range of competitions, where I have tackled complex data science challenges and honed my problem-solving skills. Through these experiences, I have not only contributed to the global data science community but have also consistently improved my data analysis and machine learning capabilities." If you're interested, **Kaggle** profile at: https://www.kaggle.com/shivam17818

## IDE & Programming Languages

- Visual Studio, PyCharm, Ananconda, Jupyter Notebook, Spyder
- Python, C

## EDUCATION

**Bachelor of Technology**

Computer Science and Engineering UP, India – 2021

## HACKATHONS

**Google Landmark Recognition -2021**

Rank -35

## ACTIVITIES & INTREST

- Cricket
- Runner