# Sonu Dixit

83750-68648 sonudixit2k@gmail.com linkedin.com/in/sonudixit

## Experience

**IDFC First Bank, Senior Data Scientist** **Nov 24 - Dec 24**

- Document Chunking, LLM-Agent ChatBot — Prompt Engineering, Tool Use

**[24]7.ai, Data Scientist-2** **Feb 22 - Oct 24**

**LLM finetuning | *Self-Hosted LLM*** *Aug 24 - Oct 24*

- We fine-tune the LLAMA 3.1-8B-Instruct on Agent-Customer conversation data. Finetuning improves performance(tone, verbosity, personalisation).

**Intent Classification | *ChatBot*** *Dec 23 - Sept 24*

- We classify customer queries into predefined intents to enable automated handling by our Chatbot. Our trained model outperforms the existing solution, LLMs(few-shot) including GPT-3.5, and other data efficient methods like SetFit.

**Knowledge Article Recommendation(KAR) | *Information Retrieval, Distillation*** *Feb 22 - Dec 22*

- During Agent-Customer conversations, the Agents need to refer to external information sources, thus increasing the response time. We train a DistilBERT on the MSMarco dataset to proactively recommend relevant articles for the passage ranking task. We achieve comparable performance on the TREC21 and TREC20 datasets.

**Answer Recommendation, Generation | *IR, RAG*** *Dec 22 - Dec 23*

- We recommend relevant queries and answers from a Knowledge Base(KB) for a user query. We tune the model trained in the KAR task on the client-specific dataset(s), and outperform the existing method.
- To generate knowledge-grounded answers to user queries in an ongoing chat, We train an end-to-end Retrieval Augmented Generation(RAG) model on the Wizard of Wikipedia dataset.

**LLM as Co-Pilot | *RAG, In-Context Learning*** *Aug 23 - Dec 23*

- We utilize LLMs to generate contextually relevant responses during conversations. We use the retriever trained in the KAR task to append the LLM context with relevant historical information.

**Disney Star India - Asst. Manager (Data Science)** **July 19 - Jan 22**

**Content Analysis and Storyline Extraction in TV Shows | *Representation learning*** *Nov 20 - Jan 22*

- We quantified production, editing-level character attributes using text, image, and audio data, and correlated these features with viewership metrics. We used clustering on the derived features to distinguish between average and high-performing shows, providing actionable insights for content producers.
- We trained episode vectors from text summary data, grouped sequential episodes into plot, and applied summarization techniques to derive storylines in the plots. We extracted keywords to quantify social interactions to construct relationship graphs.

**Entity extraction from Semi-Structured Documents | *HMM*** *July 19 - Jan 20*

- Document consists of key-value pair, tabular and text data. We need to extract values corresponding to predefined keys. We used Hidden Markov Model(HMM) to model the inherent sequential dependency in phrases.

## Publications

- Zero-Shot Generalization using Intrinsically Motivated Compositional Emergent Protocols, R Hazra*, Sonu Dixit*, S Sen in *NAACL 2021 workshop - ViGIL*
- gComm: An environment for investigating generalization in Grounded Language Acquisition, R Hazra, Sonu Dixit in *NAACL 2021 workshop - ViGIL*

## Education

**Indian Institute of Science** **2017 – 2019**
*MTech (Artificial Intelligence)* *Overall GPA: 7.5/10*

**Gurukula Kangri University** **2012 – 2016**
*BTech (Computer Science)* *Overall percentage: 78.60*

## Projects

**Adaptive Traffic Signal Control | *Multi Agent RL, PPO*** **MTech Thesis**

- Implemented multi-agent reinforcement learning (MARL) to adjust traffic signal durations based on congestion dynamically. Algorithm - Proximal Policy Optimization(PPO) with Advantage Actor-Critic
- Simulated traffic data using PTV Vissim and demonstrated MARL's superiority over fixed-time algorithms in terms of average speed, delay, and lane occupancy. [Report]