Dhaval Taunk

+91 9713899088 | taunkdhaval08@gmail.com | LinkedIn | GitHub | Portfolio | Google scholar

Education

- **International Institute of Information Technology (IIITH)** Hyderabad, India
  MS by Research - Computer Science and Engineering; ***CGPA: 10/10*** *Aug 2021 - July 2023*
  Thesis Link*: References as Building Blocks: Investigating their Significance in Encyclopedic Text Generation*

- **Indian Institute of Information Technology, Design and Manufacturing** Jabalpur, India
  Bachelors of Technology - Computer Science and Engineering; ***CGPA: 7.2/10*** *Aug 2016 - June 2020*

Experience

- **FastCode AI**
  Research Scientist *Aug 2023 - Present*
  - **SQLWizard**: Built a text2sql framework using GPT 4 model to generate MySQL query from natural language input query. The system uses only database schema which eliminates the risk of exposing the data to these models and thereby keeping the data privacy intact.
  - **DVDB**: Created a open source dart based vector database DVDB that can store vector embeddings locally on mobile devices (android/iOS). Also implemented the search functionality to retrieve the top-k relevant results.
  - **Persona based Chatbot**: Worked on to create a chatbot that can adopt different personas based on the instruction and can act and respond accordingly. Experimented with Llama 2 & Mistral 7B models to achieve the goal.
  - **Battery life cycle prediction**: The aim of this project is to build a federated learning based system that can predict remaining battery life cycle in real-time for electric vehicles. The utilization of federated learning enables prediction models to be trained directly on individual devices, ensuring data privacy and security without the need for data to leave the local devices.

- **International Institute of Information Technology, Hyderabad**
  Research Assistant - Information Retrieval & Extraction Lab (iREL) *Jan 2022 - July 2023*
  - **About**: Pursued MS by Research under the guidance of Prof. Vasudeva Varma and Prof. Manish Gupta.
  - **Role**: Worked as Research Assistant on different problems like encyclopedic text generation for low resource languages, outline generation for encyclopedic text, question answering using commonsense reasoning etc.
  - **Work**: Published half a dozen papers including XWikiGen, GrapeQA during the tenure.
  - **Additional**: Mentored a dual-degree student in a project which generates outline of encyclopedic text from references.

- **Yes Bank**
  Data Scientist *Aug 2020 - July 2021*
  - **Industry and Sub-industry Classification**: The project aims to identify potential small scale industries as customers based on their work description available on internet and help them in growing their business by recommending relevant products. Tech Stack: Python, PyTorch, Transformers
  - **Loyalty Rewards Program**: Loyalty Rewards program aimed at awarding reward points to customers based on their transaction type and given set of rules. Tech Stack: Hadoop, PySpark

- **Jio Haptik Technologies Limited**
  Machine Learning Intern *May 2019 - Nov 2019*
  - **Project**: Built an intent detection system of chatbots by finetuning and testing several deep learning based models s like BiMPM, ABCNN, BERT, Siamese based networks, USE, ULMfit, tf-idf etc.
  - **Outcome**: Improved bot's performance by using ULMfit and tf-idf and thereby leading to a 13% (approx.) rise in accuracy and an enhanced customer experience.

- **IIT Guwahati**
  Summer Research Intern *May 2018 - July 2018*
  - **Project**: Implemented Gender Classification by using deep neural networks in live video streaming by training 3 different models on image, audio and video dataset.
  - **Outcome**: The achieved accuracy of models is 87.4%, 98.7% and 68.4% for the image, audio files and video files respectively.

Publications

1. *XWikiGen: Cross-Lingual Summarization for Encyclopedic Text Generation in Low Resource Languages*
   Dhaval Taunk, Shivprasad Sagare, Anupam Patil, Shivansh Subramanian, Manish Gupta, and Vasudeva Varma.
   In Proceedings of The ACM Web Conference 2023, WWW '23.

2. *GrapeQA: GRaph Augmentation and Pruning to Enhance Question-Answering*
   ***Dhaval Taunk**, *Lakshya Khanna, *Siri Venkata Pavan Kumar Kandru, Vasudeva Varma, Charu Sharma, and Makarand Tapaswi
   In Companion Proceedings of The ACM Web Conference 2023, WWW '23 Companion

3. *XOutlineGen: Cross-lingual Outline Generation for Encyclopedic Text in Low Resource Languages*
   Shivansh Subramanian, **Dhaval Taunk**, Manish Gupta and Vasudeva Varma
   In Wiki Workshop 2023

4. *Summarizing Indian Languages using Multilingual Transformers based Models*
   Dhaval Taunk and Vasudeva Varma
   In Forum for Information Retrieval Evaluation, 2022

5. *IIIT-MLNS at SemEval-2022 Task 8: Siamese Architecture for Modeling Multilingual News Similarity*
   Sagar Joshi, **Dhaval Taunk**, and Vasudeva Varma
   In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)

6. *Profiling irony and stereotype spreaders on Twitter based on term frequency in tweets*
   Dhaval Taunk, Sagar Joshi, and Vasudeva Varma
   In Conference and Labs of the Evaluation Forum (CLEF) 2022

Skills Summary

- **Skills**: Deep Learning, Machine Learning, Natural Language Processing, Graph Neural Networks, Algorithms
- **Languages & Frameworks**: Python, PyTorch, Tensorflow-2.0, Keras, Scikit-Learn, MySQL
- **Tools**: Git, VS Code, Jupyter Lab

Projects

- **Wikipedia Search Engine**: The aim of this project was to build a search engine from scratch. The process involved creating a index and then implement search functionality on top of index to retrieve relevant results.
  - **Index Creation**: English Wikipedia dump of size 84GB was used to create index by removing stop-words, stemming the words, removing the words longer than a certain length etc. The created index had a size of 19GB.
  - **Search Functionality**: The search functionality was implemented using TF-IDF based ranking mechanism. Tech: Python, XML, NLTK, PyStemmer.
- **Question Answering using CommonSense Reasoning**: Proposed modifications (PEGA and CANP) on the model proposed by QAGNN for the task of Common Sense Question-Answering on datasets CSQA, OBQA, MedQA which involves training language models and graph neural networks simultaneously. The overall performance of the system improved for OBQA and MedQA datasets and obtained a comparable performance on CSQA dataset.
  - **Prominent Entities for Graph Augmentation (PEGA)**: Graph augmentation works by extracting noun phrase chunks c from the question and answer pair $[q; a_o]$.
  - **QA Context-Aware Node Pruning (CANP)**: It aims to remove the less relevant nodes from the WG. Our intuition is that some Extra nodes (i.e. 2-hop neighbors from the KG which do not match the QA text) may be less relevant to the QA as compared to the Question / Answer entity nodes.
- **Transliteration**: A Seq2Seq transliteration pipeline was build in this project to convert text from native Indian language to corresponding roman script. Google's Dakshina dataset was used train the pipeline and Hindi domain was used for training and testing of model. Tech: Python, Tensorflow-2.0, LSTM, GRU
- **Text Segmentation in images using Auto-encoders**: This project aimed to create a system that can perform text segmentation in images using Auto-encoders. For training purpose, KAIST Text Scene dataset was used. Tech: Python, Keras, OpenCV.
- **Salient Object Detection**: The objective of this project is to perform Salient Object Detection by implementing a paper called *Deep Embedding Features for Salient Object Detection*.

Honors and Awards

- Achieved 2^{nd} position in the shared task Indian Language Summarization organized in FIRE 2022.
- Secured a rank of 288 out of 6871 candidates in Capgemini Tech Challenge (Data Science) 2018.
- Awarded Meritorious student incentive on scoring above 85% in class 12^{th} by Madhya Pradesh Govt.

Volunteer Experience

- **DVDB**: Created an open-source Dart based on-device and privacy preserving vector database called *DVDB*
- **Writer @ AnalyticsVidhya/Medium**: Wrote technical articles related to machine learning, deep learning etc. fields for Analytics Vidhya publication on *Medium*
- **Contributor @ HuggingFace's Transformers**: Contributed 2 Community notebooks in HuggingFace's *Transformers* repository related to multi-class and multi-label text classification
- **Teaching Assistant for IRE course**: Worked as a teaching assistant for the course Information Retrieval & Extraction (IRE) in Monsoon 2022 semester @ IIIT Hyderabad and mentored 8 students in their course project.