Onkar Litake

Email : olitake@ucsd.edu
Linkedin: https://www.linkedin.com/in/onkar-litake/
Google Scholar
Github: https://github.com/Onkar-2803

EDUCATION

University Of California, San Diego **GPA: 4/4**
Master of Science - Computer Science and Engineering; Sep 2022 - Mar 2024
***Courses:** Statistical NLP, Advanced Text Mining, Probabilistic Reasoning & Learning, Recommender Systems & Web Mining*

Pune Institute of Computer Technology *Pune, India*
BE - Computer Engineering, Honours Degree in AI/ML; **GPA: 9.62/10** July 2018 - June 2022

SKILLS

Programming: Python, C++, JavaScript, Bash, HTML, CSS, TensorFlow, PyTorch, Keras, Data Visualization, GIT, MongoDB, SQL, AWS, EC2, S3, Azure, Tableau, Java, RESTful

RESEARCH EXPERIENCE

ML Research Intern
UC San Diego Health *June 2023 – Present*

- Leveraged Retrieval-Augmented Generation (RAG) and ReAct Prompting to architect a Language Model (LLM) for generating articles and summaries on emerging medical research topics.
- Developed a Language Model for CAD prediction and data extraction, incorporating advanced model interpretation techniques with the Shap tool. Utilized the tool's outputs to create concise summaries of extracted data, achieving an **84%** BERT similarity score in comparison with physician-generated summaries.
- Developed a machine learning model using GPT to identify social determinants of health (SDoH). The dataset included GPT-generated sentences for specific SDoH, like 'Homelessness,' injected into the i2b2 dataset. To mitigate model reliance on keywords, negative sentences were intentionally included, incorporating negation. Achieved ROC-AUC score of **91.2%**

Student Researcher
University of California San Diego *Sep 2022 – June 2023*

- Introduced a data-reweighting-based multi-level optimization framework for domain adaptive paraphrasing in text augmentation, employing GPT and BART. This approach significantly elevated the F1 score for LONG Covid Text classification by **17%**. The work has been accepted for publication in the Scientific Reports Journal.
- Developed IndiText Boost, a text augmentation framework tailored for low resource Indian languages. Implemented techniques including EDA, Back Translation, Paraphrasing, and Text Generation, resulting in a notable F-1 score increase of approximately **41%** for certain languages in a text classification task.
- Generated the most extensive dataset for Question-Answering in Hindi and Marathi by translating the SQuAD 2.0 dataset into these languages. Achieved an Exact Match of **48%** and a Rouge-L score of **0.66**.

Research Assistant
Pune Institute of Computer Technology *July 2020 – June 2022*

- Created the first public major gold standard named entity recognition dataset in Marathi, consisting of 25,000 sentences categorized into 8 entity classes. Developed an NER model with an F1 score of **86.80%** and an accuracy of **97.15%**.
- Achieved top ranks in multiple **A* conference** workshops across tasks such as Machine Translation, Hate Detection, Emotion Analysis, and Document Summarisation.
- **Published in workshops at ACL, EMNLP, COLING, AACL, WMT 22 (EMNLP 22), and LREC.**

SELECTED PUBLICATIONS

- **L3Cube-MahaNER: A Marathi Named Entity Recognition Dataset and BERT models**
  The International Conference on Language Resources and Evaluation (Dataset)
- **Neural Machine Translation On Dravidian Languages**
  Workshop at ACL 2022

- **Mono vs Multilingual BERT: A Case Study in Hindi and Marathi Named Entity Recognition**
  International conference on Recent Trends in Machine Learning, IOT, Smart Cities & Applications
- **Abstractive Approaches To Multidocument Summarization Of Medical Literature Reviews**
  Workshop at COLING 22
- **Unsupervised and Very-Low Resource Supervised Translation on German and Sorbian Variant Languages**
  WMT 22(EMNLP 22)
- **Improving long COVID-related text classification: a novel end-to-end domain-adaptive paraphrasing framework**
  Scientific Journal in Nature
- **Breaking Language Barriers: A Question-Answering Dataset for Hindi and Marathi**
  In review

Internship Experience

Research Associate *Singapore(remote)*
d.Kraft(IIIT-D) May 2021 – Sep 20211

- Designed & implemented a closed domain Question and Answering(QnA) model using ALBERT in conjunction with a Deep Retriever to assist students on the e-learning platform.
- Implemented Deep Retriever and voice-to-voice sub system using AWS and Azure.
- Developed a system for captioning & voice-over of videos from English to multiple Asian languages using AWS and Azure.
- Designed and developed a student-oriented chatbot using the RASA framework, enabling seamless and accurate responses to student queries.
- Coordinated the technical team being the first recruit of the startup.

Machine Learning Intern *Pune, India*
UST Global June 2021 – September 2021

- Developed a model to establish hierarchy of bugs encountered after extracting data from HSDES (Intel Tool).
- Employed classical supervised machine learning algorithms in conjunction with a shallow neural network architecture to effectively classify the severity of encountered bugs, facilitating accurate bug prioritization and resolution.