Lokesh Madasu

9133832696 Website maadasulokesh@gmail.com GitHub LinkedIn

Education

International Institute of Information Technology Hyderabad, India
Master of Science by Research in Computer Science - CGPA: 8.83 *2021 - ongoing*

Rajiv Gandhi University of Knowledge Technologies Nuzvid, India
Bachelor of Technology in Computer Science and Engineering - CGPA: 8.43 *2015 - 2019*

Experience

IIIT Hyderabad Hyderabad, India
Research Assistant under Dr. Manish Shrivastava - NLP Lab *Dec 2021 - ongoing*

- Working on text generation problems, creating high-quality datasets using crowdsourcing, with a primary focus on low-resource Indian languages. Exploring the performance of state-of-the-art pre-trained models on Indic language datasets, specifically in tasks such as summarization and headline generation.

IIIT Hyderabad Hyderabad, India
Research Intern under Dr. Manish Shrivastava - NLP Lab *May 2020 - Dec 2021*

- Developed site-specific web scrapers to collect high-quality data from a variety of news websites in 8 Indian languages, resulting in 4.8 million news article-headline pairs for headline generation task.
- Created the first largest human-annotated Abstractive Summarization dataset for Telugu.

ChatGen.ai Mumbai, India
Data Engineer Intern *Oct 2017 - Dec 2017*

- My primary focus was on tasks such as data gathering, cleaning, and extracting valuable information from various web sources.
- Also, I worked on conversational chatbots, using technologies such as Wit.ai and Recast.ai.

Research Projects

A Semantic Textual Relatedness Benchmark August 2023 - Current

- Collaborating with researchers from NRC-Canada, Cardiff University to develop a benchmark dataset for the Semantic Relatedness task, focusing on the Telugu, Hindi, and Marathi languages.
- The project involved creating 4100 sentence pairs, each demonstrating varying degrees of relatedness.

News Headline Generation for 8 Indic languages May 2022 - Aug 2023

- Created an extensive multilingual dataset comprising 3.4 million article-headline pairs for headline generation task.
- Experimented with state-of-the-art pre-trained models and achieved an average R-L score of 31.43 across all eight languages.

Relevance-based News Headline Classification Feb 2022 - October 2023

- Created the largest Telugu news headline classification dataset, comprising 26,178 article-headline pairs.
- Given a news article and headline as input, the model classifies the headline into three classes based on its relevance to the article.

Measure Text Fluency Feb 2022 - July 2022

- Implemented a language model based metric to measure the fluency in 9 Indian languages, and Sinhala.
- The results of baseline models show a strong correlation with human judgments.

Publications

- Lokesh Madasu, Gopichand Kanumolu, Nirmal Surange, Manish Shrivastava. *Mukhyansh: A Headline Generation Dataset for Indic Languages.* 37th Pacific Asia Conference on Language, Information and Computation, PACLIC-2023 (Accepted as a long paper, to be published).
- Gopichand Kanumolu, Lokesh Madasu, Pavan Baswani, Ananya Mukherjee, Manish Shrivastava. *Unsupervised Approach to Evaluate Sentence-Level Fluency: Do We Really Need Reference?.* South East Asian Language Processing Workshop, AACL-2023 (**Best Paper Award**, to be published).
- Gopichand Kanumolu, Lokesh Madasu, Nirmal Surange, Manish Shrivastava. *TeClass: A Human-Annotated Relevance-based Headline Classification and Generation Dataset for Telugu.* The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024 (Under review).

Course Projects

Abstractive Summarization for Telugu Python, Keras, Transformers

- Given a news article as input, the model generates a concise summary. Implemented various state-of-the-art transformer models on TeSum dataset and achieved ROUGE-L score of 34.0.

CodeMixed Machine Translation Python, PyTorch, Transformers

- Implemented a code-mixed machine translation system that translates English sentences into Hinglish sentences(a combination of words from Hindi and English).

Clickbait Intensity & Style Analysis Python, Keras, Transformers

- The goal is to predict and reduce clickbait intensity in headlines. Various regression algorithms and pre-trained models were used to predict intensity, along with a paraphrase model for reduction.

Wikipedia Search Engine Python, XML, Stemmer, Streamlit

- Developed a scalable search engine for a 90GB English Wikipedia dump, utilizing inverted indexing, searching and relevance ranking. Achieved query response times of under 10 seconds, handling both plain and field queries, while providing accurate and timely results in the form of relevant Wikipedia page titles.

Skills

- **Languages**: Python, C
- **Databases**: SQL, MongoDB
- **Tools & Frameworks**: PyTorch, Keras, Scikit-Learn, Huggingface Transformers

Relevant Courses

- Statistical Methods in AI, Introduction to NLP, Advanced NLP, Information Retrieval & Extraction, Data Analytics

Volunteer Experience

- Mentored over 400+ student interns in the NLP course. Designed and assessed quizzes, assignments, and projects, as well as resolved queries related to assignments, projects, and NLP concepts.
- Volunteered at the Advanced Summer School on NLP (IASNLP-2022) organized by Language Technologies Research Center, IIIT-Hyderabad.

Honors And Awards

- Won the Best Paper Award for my paper titled **Unsupervised Approach to Evaluate Sentence-Level Fluency: Do We Really Need Reference?** at IJCNLP-AACL 2023 workshop.

Reference

Dr. Manish Shrivastava (Assistant Professor, MT-NLP Lab, IIIT-Hyderabad)
m.shrivastava@iiit.ac.in