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 pretrained 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, and then storing it in MongoDB databases for various analytical purposes. Additionally, I worked on conversational chatbots, using technologies such as Wit.ai and Recast.ai.

Research Projects

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

- Created "Mukhyansh", an extensive multilingual dataset of 3.4 million article-headline pairs, tailored for Indian language headline generation task. Evaluated the state-of-the-art text generation models performance on Mukhyansh, achieving a remarkable average ROUGE-L score of 31.43 across all 8 languages, surpassing all other headline generation models.

Relevance-based News Headline Classification Feb 2022 - October 2023

- Created first-ever human-annotated Telugu news headline classification dataset, containing 78,534 annotations across 26,178 article-headline pairs. The task involves categorizing news headlines based on their relevance to the corresponding news articles into one of the 3 classes: Highly Related, Moderately Related, and Least Related. Experimented with various BERT-based models, and achieved an impressive F1 weighted average of 0.63 and an F1 macro score of 0.64.

Measure Text Fluency Feb 2022 - July 2022

- Implemented a language model based reference free metric to assess the fluency of machine-generated text in 9 Indian languages, and Sinhala. Introduced a human-annotated benchmark test dataset of 5K sentences, and the results of baseline models exhibit a strong correlation with human judgments.

A Semantic Textual Relatedness Benchmark August 2023 - Current

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

Publications

1. 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).
2. 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).
3. 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).

Skills

- **Languages**: Python, C, JavaScript, HTML, CSS
- **Databases**: SQL, MongoDB
- **Tools & Frameworks**: PyTorch, Keras, Scikit-Learn, Huggingface Transformers
- **Development**: Git, VS Code

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 the 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.

Relevant Courses

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

Volunteer Experience

- Mentored 400+ student interns in the NLP Course. Designed and assessed quizzes, assignments, and projects. Resolved queries related to assignments, projects and NLP concepts.
- Volunteered in 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