Sumanth Doddapaneni

doddapaneni.sumanth@gmail.com | sumanthd17.github.io | GitHub | LinkedIn

Education

Indian Institute of Technology (IIT), Madras — Jan. 2022 – Ongoing
Ph.D, Computer Science, Advisor - Dr. Mitesh M. Khapra

Indian Institute of Technology (IIT), Madras — Jan. 2021 – Dec 2021
External Student, Computer Science — *CGPA: 8.5/10*

Indian Institute of Information Technology (IIIT), Sri City — Aug. 2016 – May 2020
B.Tech. Electronics and Communications — *CGPA: 7.91/10*

Experience

AI Resident — Oct 2021 – Present
AI4Bharat, IIT Madras — *Chennai, Tamil Nadu*

- Working on multilingual language models
- Advised by Dr. Mitesh Khapra, Dr. Pratyush Kumar and Dr. Anoop Kunchukuttan

Post-Baccalaureate Fellow — Oct 2020 – Oct 2021
Robert Bosch Centre for Data Science & AI, IIT Madras — *Chennai, Tamil Nadu*

- Working on multilingual NLP
- Co-led the effort on Samanantar Project, SOTA NMT models on 11 Indian languages
- Worked on IndicWav2Vec Project, SOTA ASR models for 9 Indian Languages

Research Intern — Jul 2020 – Oct 2020
Swiggy Applied Research — *Bangalore, Karnataka*

- Worked on NER for unstructured chat data.
- Used weak supervision based approaches for NER detection
- Created internal tool for data annotation

Data Engineer Intern — Jan 2020 – Jul 2020
Cogitech Solutions Pvt. Ltd. — *Pune, Maharastra*

- Trained and deployed RNN based models on AWS Sagemaker
- Created end-to-end workflows on Apache Airflow

Publications - Google Scholar

1. *Towards building ASR Systems for the Next Billion Users* — [paper]
   Tahir Javed, **Sumanth Doddapaneni**, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra
   Accepted at *AAAI Conference on Artificial Intelligence, 2022*

2. *Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages* — [paper]
   Gowtham Ramesh*, **Sumanth Doddapaneni***, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh Khapra
   Accepted at *Transactions of the Association for Computational Linguistics (TACL), 2022*

3. *Bitions@DravidianLangTech-EACL2021: Ensemble of Multilingual Language Models with Pseudo Labeling for offence Detection in Dravidian Languages* — [paper]
Debapriya Tula*, Prathyush Potluri*, Shreyas Ms, **Sumanth Doddapaneni**, Pranjal Sahu, Rohan Sukumaran, Parth Patwa
Accepted at *First Workshop on Speech and Language Technologies for Dravidian Languages (EACL)-2021*

* denotes equal contribution

Pre-prints

1. *Effectiveness of Mining Audio and Text Pairs from Public Data for Improving ASR Systems for Low-Resource Languages* — [pre-print]
Kaushal Santosh Bhogale, Abhigyan Raman, Tahir Javed, **Sumanth Doddapaneni**, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M Khapra

2. *A Primer on Pretrained Multilingual Language Models* — [pre-print]
Sumanth Doddapaneni*, Gowtham Ramesh*, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra
Under review at *ACM Computing Surveys*

3. *A survey in Adversarial Defences and Robustness in NLP* — [pre-print]
Shreya Goyal, **Sumanth Doddapaneni**, Mitesh M. Khapra, Balaraman Ravindran
Under review at *ACM Computing Surveys*

Honours

- Received Post Baccalaureate Fellowship from Robert Bosch Centre for Data Science and AI
- Ranked 1st, 4th & 5th globally for Malayalam, Tamil & Kannada, respectively on Offensive language detection in Dravidian languages shared task at Speech and Language Technologies for Dravidian Languages Workshop, collocated at EACL 2021
- Finished 19th out of 167 teams with a score of 0.980 (highest score: 0.986) on the Combating Online Hostile Posts in Regional Languages during Emergency Situation workshop, collocated at AAAI 2021

Projects

XLingNLP | *Python, PyTorch, HuggingFace* Dec'20 - Ongoing
- Built a toolkit for training and evaluation of cross-lingual models
- Supports training of NMT models, LMs, fine-tuning & evaluation on XTREME benchmark

IndicWav2Vec | *Python, PyTorch, fairseq* | [code] Jun'21 - Nov'21
- IndicWav2Vec is a multilingual speech model pretrained on 40 Indian languages
- Model achieves SOTA performance on 9 languages across 3 benchmarks
- Trained and released models, training and evaluation scripts
- Paper accepted at AAAI 2022

IndinTrans | *Python, PyTorch, fairseq* | [code] Mar'21 - Oct'21
- IndicTrans is multilingual transformer trained on Samanantar dataset. Achieves SOTA performance compared to all publicly available benchmarks on 11 languages
- Trained and released models, training and evaluation scripts
- Paper accepted at TACL 2021

Service

- Program Committee at Multilingual Representation Learning (MRL) Workshop, EMNLP 2021
- Volunteer at ICML 2021, NeurIPS 2021
- Contributed to HuggingFace Datasets by adding public datasets and code