Sanidhya Vijayvargiya

Email: sanidhyavijay903@gmail.com
Github: sani903
Mobile: +91-843-3986-606
Google Scholar
LinkedIn

Education

- **Birla Institute of Technology & Science, Pilani** — India
  Bachelor of Technology - Computer Science Engineering; GPA: 9.02 — *Nov 2020 - May 2024(expected)*
  Relevant Courses: *Natural Language Processing, Data Mining, Reinforcement Learning, Information Retrieval*

Awards

- MITACS Globalink Research Internship '23 on use of NLP to generate knowledge graphs for software engineering artifacts(scholarship amount of approx. 15,000 CAD)
- Best Paper Award ICCSA 2022 for paper on Software Sentiment Analysis
- Second Runner's Up in paper presentation at CEReS 2022 - Computing and Electronics Research Summit

Experience

- **Undergraduate Research Assistant in Natural Language Processing** — Dept. CSIS, BITS Pilani
  Mentor: Dr. Lov Kumar, Assistant Professor, BITS Pilani — *Aug 2021 - Present*
  - **Research output**: Worked on 7 research papers, out of which 5 are accepted/published and 2 are under review.
  - **Set-up**: Performed literature review to find scope for improvement in the area. Built research framework with the help of prof and selected ML/NLP techniques to use.
  - **Programming tasks performed**: Pre-processed textual datasets related to software engineering and Covid-19 using stemming, tokenization. Coded the pipeline to set up various models to be compared. Monitored the models' training for efficiency.
  - **Presentation of work**: Wrote entire research paper and drew observations from the performances. Used statistical testing to validate claims. Updated prof on each step. Presented the work at conferences.
- **Machine Learning Intern in Project Udaan** — Dept. of CSE, IIT Bombay
  Mentor: Prof. Ganesh Ramakrishnan, Institute Chair Professor, IIT Bombay — *June 2022 - Present*
  - **Impact**: Project Udaan is an end-to-end machine translation ecosystem to break language barrier in education. English material is translated to local Indian languages.
  - **Role**: Working on improving post-editing tool after machine translation, and assisting with improving translation quality.
  - **Programming tasks performed**: Created trie-based find and replace algorithm from scratch, incorporating all features of regex search as searching and replacing words in translations of entire textbooks was very slow with regex search. Helped with improvements to the tool and bug fixes.
- **Data Science Intern** — New Delhi, India
  Elucidata Data Consulting Pvt. Ltd. — *May 2022 - July 2022*
  - **Role**: Processing bioinformatics data using ML techniques and manually curating cell types
  - **Programming tasks performed**: I took single cell RNA sequencing datasets from publications, selected the highly variable genes using feature selection, and removed cells which were dead using mitochondrial count, regressed out batch and sample effects from the dataset. I used dimensionality reduction to help identify clusters using leiden clustering. I then annotated the clusters found to their cell type using marker genes that were used by authors of the publication. I used U-maps, stacked violin plots, and dot plots for visualization. I stored the results using an h5ad file.
- **Undergraduate Research Assistant in Natural Language Processing** — Dept. CSIS, BITS Pilani
  Mentor: Prof. Aruna Malapati and Dr. Manik Gupta, BITS Pilani — *September 2022 - Present*
  - **Role**: Working on a project on extractive summarization of Stack Overflow posts by incorporating metadata. Simultaneously working on project on Empathic Conversational Agents.

Projects

- **Incorporating BERT into Neural Machine Translation**
  Modified re-implementation of paper published in ICLR 2020 used to translate English to Dutch
  - Uses custom tokenizer derived from Wordpiece tokenizer
  - Encoder and Decoder with multi-head attention mechanism and positional encoding for the embeddings built
  - Encoder uses a pre-trained BERT layer whereas the Decoder is custom built. The output from the Decoder is fed to a linear layer which generates the final output
  - **Project Repository**
- **Metadata-aware Summarization of Community QA platforms**
  Abstractive Summarization of various answers to questions on CQA platforms
  - Entity Pyramid sentence selection from the various answers.
  - Clustering of semantically similar sentences to incorporate different viewpoints in summary.

- Using answer metadata to select best sentences from clusters.
- BART, TD5, and Longformers transformers used to generate summary from selected sentences from each cluster. The encoder of these transformers is fine-tuned, and the final layer of the encoder incorporates metadata of the question.
- Results are not available yet as models are being trained currently.

- **Fast Find and Replace algorithm**
State-of-the-art Trie-based algorithm which incorporates all features of regex search for very large documents
  - Prefix search and multi-word search queries are supported in the trie.
  - Replace feature provided for prefixes and the remaining trie after the replaced words is merged with the trie of the new word.
  - Replace feature also works for multiple words and the trie accounts for the changes in indices of words without updating the entire trie again.
  - Can be integrated with a change-tracking tool which returns vector of words with their indices in latest traversal. Upon calling the update function, changes are detected and accordingly trie is updated to account for manual changes made without using in-built trie functions and then be deployed in a formal setting.
  - To improve efficiency, indices are stored in trie only at time of insertion of word. Change in indices of words is calculated instead.

Publications

- **COVID-19 Article Classification using Word-Embedding and ELM with Various Kernels**: First author for paper published at 36th International Conference on Advanced Information Networking and Applications AINA 2022: Core B.
- **Software Requirements Classification using Deep-learning Approach with Various Hidden Layers**: First author for paper accepted at 17th Conference On Computer Science and Intelligence Systems FedCSIS 2022: Core B.
- **Software Functional Requirements Classification using Ensemble Learning**: First author for paper published at The 22nd International Conference on Computational Science and Its Applications ICCSA 2022.
- **Software Sentiment Analysis using Machine Learning with Different Word-Embedding**: Second author for paper published at The 22nd International Conference on Computational Science and Its Applications ICCSA 2022.
- **COVID-19 Article Classification using Word-Embedding and Different Variants of Deep-Learning Approach**: First author for paper published at Fifth International Conference on Applied Informatics ICAI 2022.
- **Software Engineering Comments Sentiment Analysis using LSTM with Various Padding Sizes**: First author for paper accepted for review at 18th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2023): Core B
- **Empirical Analysis for Investigating the Effect of Machine Learning Techniques on Malware Prediction**: First author for paper accepted for review at 18th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2023): Core B

Skills Summary

- **Languages**: Python, Bash, TensorFlow, Keras, Pytorch, C++, MySQL, SQLite, JAVA, HTML, CSS, Matlab, Flask, QT
- **Libraries**: Scikit, NLTK, SpaCy, Pandas, NumPy, OpenCV, Matplotlib, Seaborn
- **ML**: SVM, Neural Networks(DNN, CNN, LSTM), kNN, SMOTE, ANOVA, PCA, Ensemble Learning, Statistical testing(Friedman test), Transformers, Encoder-Decoder with Attention, Genetic Algorithm
- **NLP**: Word Embedding, BERT, Sentiment Analysis, Text Classification, Neural Machine Translation, Abstractive Summarization, Extractive Summarization, WordPiece Tokenization
- **Soft Skills**: Leadership, Event Management, Report Writing, Public Speaking, Time Management

Extracurricular Activity

- **Thrust Vector Control Team**
Students for the Exploration and Development of Space
  - Using Reinforcement Learning in gimbal in rockets for trajectory correction. Currently implementing the DDPG algorithm for the same.

- **Co-Founder**
Students for Development of Thinking Computer Systems
  - Weekly discussions and presentations of important publications in NLP, CV, and DL to stay up-to-date with the latest developments and approaches in ML fields. Re-implementations of these papers are also attempted.
  - The goal is to improve the research culture on campus and expose my peers and me to the latest technologies.

- **Machine Learning core team**
ACM, college chapter
  - Organized multiple workshops and designed ML hackathons to promote research and development culture on campus.

- **Machine Learning Team**
Google Developer Student Clubs, college chapter