BHAVISH PAHWA

f20180168@hyderabad.bits-pilani.ac.in
linkedin.com/in/bhavishpahwa/
github.com/bp-high
Personal Website:- bp-high.github.io/

EDUCATION

Bachelor of Engineering | *Computer Science* — Aug. 2018 – June 2022
BITS PILANI HYDERABAD CAMPUS — Hyderabad, India

WORK EXPERIENCE

SDE-1(ML) — July 2022 -
MindTickle! — Remote

- Part of the pan organisation R&D and Central ML team known as COEML(Centre of Excellence for ML)
- Solving problems related to analytics and insights from sales meetings using audio and textual based solutions.
- Working on defining technical metrics for various business problems to better align and compare different NLG based approaches for automatic summarization and paraphrasing to generate byte-sized insights
- Improved the data labelling process by reducing turnaround time for annotation by creating annotator friendly labelling tools and streamlining the data pipelines. Recorded an improvement of 65% on average time spent by annotator by comparing equivalent tasks.
- Compared the use of GPT3(using OpenAI API) and in-house models for solving various business problems. Worked on developing a library for constrained text generation and prevention of hallucination in several models and also controlling and filtering the outputs generated by GPT3.

SDE Intern — Jan 2022 - Jun 2022
Amazon India — Remote

- Working in the **Amazon Academy Organisation**. My team works towards Monetization and Upper Funnel

Unicorn Intern(Data Science Intern) — Jun 2021 – Aug 2021
DBS Bank (Development Bank of Singapore) — Hyderabad , Telangana

- **Tech Stack of the engineering Division : Java,Spring framework , MariaDB ,Front End Components/ UI /UX using JavaScript frameworks like React, AngularJS , HuggingFace(Transformers)**
- Worked closely with the Data science and NLP team at the Credit & Operational Risk Department at DAH2 centre of DBS Bank , Singapore
- Worked on building a Credit & Operational Risk digest repository that scours the different confidential and public financial articles/documents that DBS procures regarding situations which can affect Credit & Operational Risk
- Worked on extracting relevant information from the Articles by using proper data cleaning and processing techniques and using Abstractive and Extractive summarization methods
- Improved the summarization pipeline by 40% by implementing a novel extractive summarization solution using **LexRank algorithm combined with Sentence Transformer based Sentence embeddings**. Used **ROGUE** as the evaluation metric for summarization
- Implemented an Unsupervised Keyword/Key phrase Extraction pipeline inspired by **EmbedRank (https://arxiv.org/pdf/1801.04470.pdf)** using **Spacy based pipeline models for candidate selection combined with sentence embeddings and maximal marginal relevance (MMR) for candidate ranking to extract top n keywords/key phrases**
- Implemented an entity extraction and tagging/linking based pipeline using spacy transformer pipeline for extraction of entities combined with **vocabulary based approach** to tag monetary entities regarding their context as to which domain of credit risk they belong to and similarly tagged organizational entities with their type of organization (like bank, institute , corporation)

PUBLICATIONS

BpHigh@TamilNLP-ACL2022: Effects of Data Augmentation on Indic-Transformer based classifier for Abusive Comments Detection in Tamil
(Accept(poster))(First Author)(ACL 2022 Workshop DravidianLangTech)
Implementation Code(currently in the process of open sourcing):-**GitHub Repo**
Paper Link:- **ACL Anthology Paper Link**

Open Source Contributions

- **Hugging Face** - PR/Commit **#3972**
- **Keras-IO** - PR/Commit **#940**
- **AcademicPages** - PR/Commit **#925**
- **Hugging Face Keras-IO sprint** - Model/Spaces Link:- **Node2Vec_MovieLens**

Projects

Multilingual Abusive Comment identification Oct 2021- Dec 2021

Kaggle Competition by Moj/ShareChat Hyderabad , Telangana

- Working on multilingual data set provided by Moj team during their Open Kaggle competition for improving systems of Abuse detection in multilingual/Indic setting
- Using pre-processing techniques relevant to Indic languages for proper use of the data set. Used Indic NLP library for text normalization
- Converted emojis in the comments data set to text and then translated to the relevant language of the specific comment for better representation of sentiment in the comment
- Using **MURIL Bert multilingual model** for fine-tuning and making a relevant network using model as embedding layer and relevant classifier layers for getting predictions
- Using Stratified K-fold cross validation method at train time.
- Using Data-Centric approaches to improve model performance, using **data augmentation in NLP** to improve model performance

Technical Domain Identification Oct 2021-Dec 2021

Under Dr Ayan Das BITS Hyderabad

- Develop system/s that automatically identify the technical domain of a given text (a small passage) in specified Langauge (English, Bangla, Gujarati, Hindi, Malayalam, Marathi, Tamil, Telugu). Such a piece of text provides information about specific Coarse grained technical domains like Computer Science, Physics, Life Science, Law etc or the Fine grained subdomains for Computer Science domain, it can be of Operating System, Computer Network, Database etc.
- Using **MURIL Bert multilingual model** with layers frozen and making a relevant network using model as embedding layer and relevant classifier layers for getting predictions
- Using **Gradio framework** for developing the system
- Based on the Technical Domain Identification shared task at ICON 2020
- Project Report :- **Project Report Drive Link**

Reflaktor (Hostel management web app) Jan 2021 - Apr 2021

BITS Hyderabad Hyderabad , Telangana

- Worked in a team of 5, building a hostel management based web app product for college administrative functioning for the purpose of easing gradual reopening of college(the product idea was facilitated and designed before wave 2 hit the country)
- Followed the full software development life cycle from evaluation of AS-IS Work System and problem statements to requirements gathering to dash boarding and designing a scalable system design for further development of the project
- Worked as a Product owner + developer and maintained proper sprint planning approach with relevant discussions with development team as well as with the other major stakeholders
- Also worked on wire frame prototyping for UI/UX using FIGMA interface design tool
- GitHub repo:- **Github Repo Link and deployed product link**
- Product Report :- **Project Report Drive Link**
- Got around 25-30 users during initial release and demo phase but later couldn't release for general usage in campus due to wave 2 and further lockdown and Online Semester extension notices
- Used Pivotal Tracker for sprint planning and progress tracking and MERN stack as the tech stack for development and Heroku for deployment

Honors and Awards

NTSE Scholarship 2016

Recognition for being in top 1000 students in NTSE exam in India

Volunteer Experience

Student Volunteer at **ACL Conference 2022**

Skills

Programming: Python (NumPy, SciPy, Matplotlib, Pandas, Tensorflow 2.0, Pytorch, StreamLit, Scrapy Framework, Hugging Face Transformers), Java,Shell, C++(STL), Oracle Pl/SQL, Flask, MongoDB , Spring Framework, React JS, React Native, Javascript, JSP
Document Creation: Microsoft Office Suite, LaTex, Markdown