Bhavish Pahwa

bhavishpahwa@gmail.com
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.**
- Conducted research on how to convert GPT-3.5 generated text to gender-neutral text and **developed an in-house system to convert gendered text to gender-neutral text** (Compliant with the EU's definition of **Gender-Neutral Language**)
- Devised a domain-centered **reference-free evaluation metric for automatic evaluation of summarization and paraphrase generation tasks**
- Deployed various in-house models and devised a **framework for model-monitoring in production environments**
- Developed and deployed a service to detect and extract **next steps and action-items from meetings leveraging meeting transcripts** by creating a **fine-tuned GPT-3.5 model (https://platform.openai.com/docs/guides/fine-tuning) and prompt engineering**
- Developed and deployed a service to create **meeting minutes of sales-meetings by leveraging prompt engineering and OpenAI GPT-3.5 model**, also **devised a ranking algorithm to get the top-10 meeting minutes to be served as key-moments of the respective meeting**. Used in-house domain-specific classifier models and in-house gender-neutral conversion system to enrich the key-moments' feature
- Developed an **internal playground for prompt-engineering for internal stakeholders(Devs+ML Engineers)**. Added robust automated tests to the prompt-engineering playground to **test for Mindtickle's Responsible AI policy as well as against prompt-injection and prompt-leaking attacks.** Extended the APE(Automatic Prompt Engineering) tool (https://sites.google.com/view/automatic-prompt-engineer) to work with gpt-3.5-turbo and deployed it in the playground
- Working on **improving search systems at Mindtickle using LLMs by leveraging hybrid search approach and re-ranker models(fine-tuned cross-encoders)**
- Single handedly spearheading the migration of all production services from **OpenAI Platform API to Azure OpenAI Service.**
- Skills: Large Language Models (LLM) · Information Retrieval · Vector Database · AWS Opensearch · Retrieval Augmented Generation(RAG) · GPT-4 · GPT-3.5 · Generative AI · Prompt Engineering · Amazon Web Services (AWS) · Data Science · Computer Science · Natural Language Processing (NLP) · Continuous Integration and Continuous Delivery (CI/CD)

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

- Working in the **Amazon Academy Organisation**. My team works towards Monetization and Upper Funnel
- Skills: Computer Science · Continuous Integration and Continuous Delivery (CI/CD) · React-Native · ReactJS

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)
- Skills: Data Science · Computer Science · Natural Language Processing (NLP)

Research Intern May 2019 – June 2019
LeadingIndia.ai Greater Noida

- Worked towards making a reinforcement learning based agent for the Pommerman challenge of Neurl IPS 2019 :-**Pommerman**
- Worked on multi agent reinforcement learning based policies and implemented different reinforcement learning algorithms so as to improve the performance of the agent.
- Successfully built an agent that performed as efficiently as SimpleAgent which was a baseline heuristic using DQN and an architecture inspired by AlphaGo and Atari papers.
- Most of our agents emerged with defensive behaviors where we tried to train them further with reward shaping to observe emergence of other behaviors.
- Medium article:-**Automated Gaming pommerman-ffa**
- arXiv Preprint:-**https://arxiv.org/ftp/arxiv/papers/1907/1907.06096.pdf**
- Further Details:-**https://www.leadingindia.ai/projectdetails/64**
- Github Code:-**LeadingIndia**

PEER-REVIEWED PUBLICATIONS

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

- We explored the effects of data augmentation techniques on the Indic-Transformer based classifier created using MURIL Transformer on the task of Abusive Comment Detection in Tamil. We observe a negative result in the case of word-level augmentation using Contextual Models(MURIL) and an improvement in performance in the case of augmentation using Non Contextual Word Embeddings(Tamil fastText).
- Our paper makes a two-fold contribution to the shared task. First, we experiment with the state-of-the-art transformer models pre-trained on Indian languages. Secondly, we show how data augmentation techniques in NLP perform in this task and how training with word-level augmented sentences affect our model accuracy.
- As we further try to speculate why our augmentation technique based on Contextual Models failed to yield a better result, we consider the reasons stated in **Longpre et al. (2020)**, which show that data augmentation techniques help improve performance on the task only when the approaches provide a language pattern that is not seen before during pretraining of the Transformer model. As both the Contextual Model for augmentation and the Indic-Transformer used to create the classifier is MURIL transformer, we cannot observe new linguistic patterns.
- Also, in **Kobayashi (2018)**, the authors observe that augmentation based on Contextual Models might not be able to remain compatible with the annotated labels of the original input and thus might harm the training process. They suggest using information from both label and context to generate word-level augmentations to control this incompatibility

2) BpHigh at SemEval-2023 Task 7: Can fine-tuned cross-encoders outperform GPT-3.5 in NLI tasks on clinical trial data?
(Accepted)(First Author)(ACL 2023, SemEval Workshop)
Implementation Code:- Coming Soon
Poster Link:- **Semeval Poster**
Paper Link:- Paper will be presented and published after ACL 2023 conference in July

- We participate in subtask 1 only and demonstrate how cross-encoder models fine-tuned on small training datasets can outperform GPT-3.5 in zero-shot settings. We fine-tune several models based on sentence transformer and cross-encoder architecture for sentence pair modeling. We also compare our models with GPT-3.5 Davinci (text-davinci-003 according to the OpenAI platform) in zero-shot and few-shot settings. We also show the effect freezing of base transformer layers has while training cross-encoders on their performance.
- We map the textual entailment task as a sentence-pair classification task, and we study the literature to find all approaches that can be used to solve this task. We also study clinical literature to grasp our dataset and understand the definition of whereas properties properly. We also try to find pretrained models already trained on some clinical or biomedical domain that can be finetuned further for our task.
- We agree on specific approaches/pretrained models and finetune our models based on those pretrained models. To get a good grasp of the performance of our models, we leverage very large language models and compare performance with GPT-3.5 Davinci state-of-the-art models in zeroshot and few-shot settings. We devise a few-shot strategy based on semantic similarity to find the top two few-shot train snippets for each test sample. We observe that finetuned cross-encoders perform well compared to zero-shot GPT-3.5, and some models also perform comparably to the few-shot GPT-3.5.
- We also observe that freezing the base transformers' layers while training cross-encoders considerably affects model performance on downstream tasks.

3) BpHigh at WASSA 2023: Using Contrastive Learning to build Sentence Transformer models for Multi-Class Emotion Classification in Code-mixed Urdu
(Accepted)(First Author)(ACL 2023, WASSA Workshop)
Implementation Code:- Coming Soon
Poster Link:- **WASSA Poster**
Paper Link:- Paper will be presented and published after ACL 2023 conference in July

- We describe our approach in this paper for the MCEC track of the shared task. We leverage the unsupervised training method using contrastive learning for developing a sentence-transformer model from MuRIL pre-trained model for romanized codemixed Urdu.
- We leverage this sentence-transformer model to build multi-class classifiers using the provided training data and the SetFit framework.
- We show how increasing the value of the hyperparameter number of iterations increases the performance of the classifiers.

Open Source Contributions

- **Hugging Face Datasets Library** - PR/Commit **#3972**
- **Keras** - PR/Commit **#940**
- **AcademicPages** - PR/Commit **#925**
- **Hugging Face Keras sprint** - Model/Spaces Link:- **Node2Vec_MovieLens**
- **llama-hub**- PR:- **#241 #285**
- **Pinecone** - PR:- **#217**

Projects

Cross Sectional Data Analysis(EDA) of UCI Computer Hardware Dataset Sep 20 - Nov 20
Exploratory Data Analysis , Econometrics Course BITS Hyderabad , Telangana

- Explored the dataset using both simple and multiple linear regression and estimating model criteria for selecting a proper functional form for the final model
- Achieved a R-squared value of 0.86(out of 1) on the final model with a p-value of 2.2e-16 indicating significance of the chosen functional form and the explanatory variables
- Performed various tests to check that none of the assumptions of Multiple Linear regression(like conditional Homoscedasticity, error normality) are violated
- **Project Report Part 1**
- **Project Report Part 2**

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, Prompt Engineering, Generative AI
Document Creation: Microsoft Office Suite, LaTex, Markdown