




Arpit Dwivedi

 github.com/arpitdwi  linkedin.com/in/arpitdwi  arpitdwived@gmail.com  +91-8976058580

EDUCATION

Indian Institute of Technology Kharagpur

2017 - 2022

B.Tech + M.Tech (Dual Degree)

CGPA: 9.2/10.0

- **Major:** Electronics and Electrical Communication Engineering
- **Minor:** Computer Science and Engineering
- **Micro:** Entrepreneurship and Innovation

Delhi Public School, Navi Mumbai

2017

All India Senior School Certificate Examination

Marks: 94.6%

Delhi Public School, Navi Mumbai

2015

All India Secondary School Examination

CGPA: 10/10

EXPERIENCE

Data and Applied Scientist, Bing Ads, Microsoft

Jun 2022 - Present

- Fine-tuned XLM-Roberta based multilingual models pretrained on ad specific data for predicting the relevance between query-ad pairs with ROC-AUC score of 91.96%, marking a 4-point improvement from the baseline model
- Employed parameter-efficient fine-tuning techniques on small language models (SLMs) to evaluate their viability as alternatives to the aforementioned teacher models
- Innovated GPT prompts to generate relevance labels resulting in improvement of over 5 AUC points compared to human labels and enabling the transition from human-labeled training data to GPT-generated labels
- Explored architectures for long document understanding such as Reformer and Longformer to facilitate comprehensive understanding of landing pages and significantly reduce training time
- Carried out knowledge distillation to train compact BERT models and DNNs for efficient online deployment
- Engineered novel features and trained models using curriculum learning strategies to enhance model robustness

Data and Applied Scientist Intern, Bing Ads, Microsoft

May 2021 - Jul 2021

- Worked on identifying spoof domains in ads using Siamese Convolutional Neural Networks
- Curated a comprehensive dataset for training using manual generation techniques and achieved a validation ROC-AUC score of 98.93% using the Triplet loss function
- Leveraged Faiss for efficient nearest-neighbor searches, enabling the rapid identification of existing spoof URLs which could subsequently be used as a dataset

Computer Vision Intern, Proof of Performance Data Services Pvt Ltd

Apr 2020 - Jun 2020

- Deployed You Only Look Once algorithm to develop a tree detection system, enabling the identification of all trees within a given image or video and classified them with an accuracy of 88% using convolutional neural networks
- Implemented non-maximum suppression and performed image processing using OpenCV to filter out the false positives post tree detection leading to an increase in precision

Natural Language Processing Intern, Glassquid.io

Oct 2019 - Jan 2020

- Built a classifier to distinguish between IT and non-IT jobs using their job descriptions with an accuracy of 93% using TF-IDF weighted Word2Vec and XGBoost
- Utilized SQL for data extraction and analysed the trends in skills and experience of users, their preferred job roles, location etc. to get valuable business insights

PAPERS AND PUBLICATIONS

- [1] Rima Hazra, Arpit Dwivedi and Animesh Mukherjee. Is this bug severe? A text-cum-graph based model for bug severity prediction. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 236-252. Springer, 2022.
- [2] Bishal Santra, Ravi Ghadia, Arpit Dwivedi, Manish Gupta and Pawan Goyal. CORAL: Contextual Response Retrievability Loss Function for Training Dialog Generation Models. *Preprint*, arXiv:2205.10558v2.

SKILLS

Relevant Coursework: Algorithms, Machine Learning, Deep Learning, Matrix Algebra, Probability and Stochastic Processes, Image Processing, Big Data Processing, Natural Language Processing

Languages and Libraries: C/C++, Python, SQL, L^AT_EX, PyTorch, Hugging Face, Numpy, Pandas, Matplotlib, Scikit-Learn, OpenCV, NetworkX, PySpark, Streamlit

Systems / Platforms: Git, Windows, Linux

PROJECTS AND COMPETITIONS

Discourse mutual information based evaluation metric Mar 2022 - May 2022

- Proposed a novel unreferenced metric for dialogue evaluation using transformer encoders pre-trained using Discourse Mutual Information based loss function that enabled the model to capture intricate relationships
- Created a response pool for scoring and outperformed strong baselines on correlation with human judgements

Multilingual news article similarity Mar 2022 - Apr 2022

- Measured the similarity between mono/cross lingual news article pairs by fine tuning encoders and stacking
- Used mBERT models for predicting six different labels such as geography, narrative, style etc. and an artificial neural network (ANN) as the meta model to get the final similarity score
- Utilized Mean Squared Error (MSE) loss during model training, resulting in a notable Pearson's Correlation Coefficient of 0.702 on the final evaluation dataset

RL based training of generative dialogue systems Jul 2021 - Feb 2022

- Trained transformer based generative dialogue systems using scores from retrieval dialogue systems instead of using traditional cross entropy loss based methods that solely rely on the ground truth response
- Used scores from retrieval models like BERT and ESIM as reward for training the model using Reinforce algorithm
- The final model demonstrated significant improvements in response quality upon evaluation in terms of diversity as well as coherence with the context as compared to the conventional generative dialogue systems

AbInBev Maverick 2.0 Hackathon Apr 2021 - May 2021

- Developed an application to recommend customized on-invoice and off-invoice discounts. Reached the grand finale among 750+ competing teams and was applauded by the panelists for outstanding approach
- Employed feature engineering and outlier detection techniques before training a combination of classification and regression models to accurately predict each component of the discount using an ensemble of gradient boosted trees

Early identification of severe bugs in Ubuntu Jan 2021 - Apr 2021

- Predicted the severity of new bugs in advance using text, graph and metadata-based features with 74% accuracy
- Constructed various networks such as a bug-bug network based on affected packages using the NetworkX library and extracted graph features like Degree, PageRank, and Clustering Coefficient
- Used Doc2Vec and SBERT for learning the text features present in the bug descriptions and comments

Imposter Detection Oct 2020 - Nov 2020

- Authenticated users on the basis of their mouse activity using features like click time, pause time, cursor velocity
- Used a supervised self organizing map (SOM) on top of an unsupervised SOM to make predictions with 83% recall

AWARDS AND ACHIEVEMENTS

- Secured All India Rank 946 in JEE Advanced 2017 among 150K+ students who appeared for the examination
- Secured All India Rank 1002 in JEE Mains 2017 among nearly 1.2M students who appeared for the examination
- Certified for being in the top 1 percentile in National Standard Examination in Physics held in November 2016
- Secured All India Rank 192 in Kishore Vaigyanik Protsahan Yojana 2015 conducted by Indian Institute of Science
- Cleared National Talent Search Examination in Class X and was one of the 775 students selected for the scholarship