Priyam Basu

Contact Number Email Linkedin Github Google Scholar

EDUCATION

Masters of Science in Computational Linguistics 2024-2026

University of Washington, Seattle, USA

TECHNICAL SKILLS

- Languages: Python, UNIX, Java, C
- Tools: PyTorch, TensorFlow, Huggingface, Keras, Scikit-Learn, Pandas, GCP, Azure, AWS, Git, Docker
- Areas: Natural Language Processing, Large Language Models, Computer Vision, Machine Learning, Agile Development

INDUSTRIAL PATENTS

- "Systems, Methods and Media for retreiving an enitity from a data table using semantic search". Priyam Basu, Zhao Jie Kok, Anbarasan Selvarasu and Abhishek Pradhan. US Patent - US20230259507A1. 2023

INDUSTRIAL/RESEARCH EXPERIENCE

Babbel, Germany | *Senior Machine Learning Engineer* Aug 2023 - Present

- Developed a novel audio-to-phoneme classification package using OpenAI Whisper model which improved the performance of pre-existing module by 50%. Added a voice activity detection filter for silent noise removal from user audios.
- Curated datasets to include diverse gender groups for bias-free training, by implementing audio feature filtering methods.
- Created a grapheme-to-phoneme conversion library for German, French, Spanish, Italian, Dutch and Turkish languages.
- Migrated existing infrastructure on phoneme recognition and grapheme-to-phoneme modules from AWS Lambda to Sagemaker using Inferentia Neuron chips. Set up training, inference and MLOps pipelines including model quantisation which improved latency by 4x and reduced developer cost of the team by 20%. Deployed the model using Cloudformation for infrastructure management and designed latency-based autoscaling policies to serve 1M users daily.
- Created an end-to-end user satisfaction determination pipeline comprising a feature-specific filtering component followed by review sentiment analysis to detect product pain-points.
- Mentored junior engineers in the team for their career transition to senior-engineer track and designed the MLE career development framework for the organization.

Automation Hero, Germany | *Applied Scientist / Machine Learning Engineer* May 2022 - Jul 2023

- Developed a Document Splitting tool which categorises individual pages of multi-page documents. Designed an unsupervised novel combined approach based pipeline consisting of an algorithmic design of multimodal clustering, OCR, multi-layer contrastive key-phrase extraction, few shot learning based document classification using SetFit contrastive learning and FAISS similarity search algorithm. Achieved an accuracy of 96.5% (evaluated on real world Insurance Claims documents) while reducing a 10 hour manual process to 25 minutes automated.
- Created a Machine Readable Zone (MRZ) detector and extractor from passports and national visas from multi-page documents as scanned copies and photographs with document noise and image object skewing. Designed a novel pipeline consisting of detector and parser components, achieving an accuracy of 93.56%.
- Developed a machine text plus handwritten text single OCR framework with ability for checkbox detection along with label mapping and associated question detection. Implemented a lined clustering algorithm and snippet slicing to avoid word detection overlap which avhieved an accuracy of 92.25%.
- Worked on other projects object detection, document denoising, offline text translation, model quantisation using Onnx and deployment on GRPC.

Taiger, Singapore | *NLP Research Engineer* Jun 2021 - Apr 2022

- Created a Table Entity Detection system for question answering over natural langauge databases including aggregation operation for advanced queries. Designed a custom table categorisation module using generating equivalent categorical table embeddings. Created a novel zero shot semantic algorithm for entity detection on financial tabular data. It outperformed Google's state-of-the-art TaPas algorithm for information retrieval with accuracy of 97.79% and was computationally inexpensive without having much negative carbon impact. Published a US Patent as lead author.

- Conducted preliminary research Neural Semantic Search Engine using semantic similarity and key phrase extraction, along with dense and sparse retrieval of documents and metadata for filtering.

Nanyang Technological University, Singapore | *NLP Researcher* Mar 2021 - Jan 2022

- Created a novel text ranking metric to pick the most diverse and semantically similar augmentations from text generation models which increased accuracy on downstream classification task upto 35%. This work was accepted at GEM Workshop at EMNLP 2023 and was done under the supervision of Dr Chng Eng Siong.
- Created a novel text augmentation algorithm using Pegasus which can be integrated with any existing NER pipeline. Performed benchmark test on GMB dataset where it was able to improve the performance of baseline NER performance from F1-0.40 to F1-0.74 without making any changes to the model architecture.
- Worked on Name entity recognition for a scalable Emergency Response System using BERT with BILOU tagging primarily focussing on Medical emergencies for automated dispatch of medical teams depending on correct assessment of type of medical emergency.

Forty4Hz, India | *Data Scientist* Jun 2020 - Jan 2021

- Created a Search Bar Recommendation system from scratch using multiple optimisation techniques for an AI- based fully automated Data Analytics software, retrieved data using MySQL databases, AWS S3 buckets using Boto3, Clickhouse and deployed using FastAPI.
- Automated creation of metadata from large size excel and csv files, retrieving them from AWS S3 buckets using Boto3, for Upload Application development, for automated upload of client specific data to Clickhouse Database, thereby reducing the same 6 hour manual process for uploading to less than 5 mins using 2 clicks and front-end integration, including server deployment using FastAPI.

PUBLICATIONS

- “Benchmarking Differential Privacy and Federated Learning for BERT Models”. Priyam Basu, Tiasa Singha Roy, Rakshit Naidu, Zumrut Muftuoglu, Sahib Singh and Fatemehsadat Mireshghallah. ICML 2021 Workshop on ML4Data
- “Privacy enabled Financial Text Classification using Differential Privacy and Federated Learning”. Priyam Basu, Tiasa Singha Roy, Rakshit Naidu and Zumrut Muftuoglu. EMNLP 2021, EcoNLP Workshop
- “RankAug: Augmented data ranking for text classification”. Priyam Basu, Tiasa Singha Roy. EMNLP 2023 GEM Workshop.
- “CyberPolice: Classification of Cyber Sexual Harassment”. Priyam Basu, Tiasa Singha Roy, SohamTiwari and Saksham Mehta. EPIA 2021, Springer LNCS
- “Multimodal Sentimental Analysis of #MeToo Tweets using Focal Loss (Grand Challenge)”. Priyam Basu, Soham Tiwari, Joseph Mohanty and Sayantan Karmakar. IEEE BigMM 2020
- “But how robust is RoBERTa actually?: A Benchmark of SOTA Transformer Networks for Sexual Harassment Detection on Twitter”. Priyam Basu, Tiasa Singha Roy and Ashima Singhal. IEEE I-SMAC 2021
- “Interpretability of Fine-grained Classification of Sadness and Depression”. Priyam Basu, Tiasa Singha Roy, Rakshit Naidu and Aman Priyanshu
- “Zero Shot Table Entity Detection using pre-trained Language models for Financial Documents”. Priyam Basu and Zhao Jie Kok
- “Text augmentation using Finite State Transducers and Text summarization approaches for Name Entity Recognition”. Priyam Basu, Kway Zin Tun and Chng Eng Siong

ACHIEVEMENTS AND HIGHLIGHTS

- Lead author of US patent during internship at Taiger.
- Published 7 research papers during undergraduate studies as the lead author with 107 citations and h-index 4.
- Youngest senior-scale engineer at the age of 23 out of 1200+ employees at Babbel.
- Co-founded a profitable Ed-Tech startup called DigiTutor which enabled peer-to-peer teaching of industrial courses between seniors and juniors of universities.