Amanpreet Singh

amanpreet.singh@stonybrook.edu | /amanpreet-singh-k
+1 631-312-2565 | /amanpreet692

Education

- **Stony Brook University** — Stony Brook, NY
  Master's in Computer Science; GPA: 3.97 — *2019 – 2021*
  - **Thesis**: Sequence Labeling for Network File System Specifications
    Advisor: Prof. Niranjan Balasubramanian
  - **Courses**: Natural Language Processing, Machine Learning, Data Science, Probability & Statistics
- **University of Mumbai** — Mumbai, India
  Bachelor of Engineering in Information Technology; First Class with Distinction (73%) — *2011 – 2015*
  - **Courses**: Data Structures & Algorithms, Artificial Intelligence, Discrete Mathematics, Databases

Publications

- **Singh, A.**, & Balasubramanian, N. (2020), "Open4Business (O4B): An Open Access Dataset for Summarizing Business Documents", *Workshop on Dataset Curation and Security - NeurIPS 2020*
- Nayak, A., Acharya, N., **Singh, A.**, Sakhapara, A., & Geleda, B.(2015), "Visualization of Mechanics Problems based on Natural Language Processing", *International Journal of Computer Applications*, 116(14)

Projects

- **NER for system specifications**: Generating meaningful representations of file system specifications by classifying token sequences as code entities. The target dataset was annotated using Brat. The pre-trained BERT and RoBERTa language models were first fine-tuned on a manually scraped corpus for domain adaptation and then on the target task.
- **Startup Acquisition Prediction**: Implementation and analysis of three multi-class ensemble models including anomaly detection, Naive Bayes and random forest on highly imbalanced data to predict whether a startup will be acquired.
- **Online Toxicity**: Multi-label classifier to detect toxicity/hate in Wikipedia Comments and transfer learning experiments with the classifier on Twitter dataset. Analyzed the results of stacked LSTM/GRU against BERT and distilBERT models.
- **Long Documents Classification**: Parsing and categorization of corporate PDFdf documents with over 10k tokens into 6 labels using techniques like Bag of words, Tf-Idf, Doc2Vec and Attention based Neural models.
- **Chess Player Ratings**: Predicting the Elo rating of a chess player from the moves sequence. Efforts involved EDA and smart feature engineering using Pandas and Matplotlib; as well as modeling with Linear Regression and Random Forest.
- **Physual**: A text to scene generation system to visualize Physics problems with StanfordCore NLP, Java3D and Blender.

Experience

- **SS&C Intralinks** — Boston, MA
  Machine Learning Engineer (NLP) — *May 2020 – Dec 2020*
  - **Abstractive Summarization**: Business document summarization system built on deep learning and REST frameworks:
    1. Curated and published a dataset of 18k open access business articles with their abstracts as summaries.
    2. Improved ROUGE score of SOTA seq2seq models like BART and T5 by more than 10 points via fine-tuning.
    3. Built a custom encoder-decoder to compress larger inputs by 50% and avoid out of memory issue during training.
    4. Adapted the fine-tuned model to ONNX quantization format reducing its size by 75% and inference time by 30%.
    5. Flask based service to return the raw abstractive summary with the salient parts of the PDF highlighted.
- **J.P. Morgan Chase & Co.** — Mumbai, India
  Senior Software Development Engineer — *Feb 2018 – Aug 2019*
  - **NLP Query Service**: An interactive system to resolve user queries that uses a model trained on the CRF classifier from StanfordCore NLP and returns the nearest possible solution from an existing knowledge base.
  - **Trader Analytics**: Introduced statistical enhancements in the core application such as absolute and percent variance, market share and standard deviation of historical stock prices to aid in trading decisions.
  - **Real-Time Pricing**: Developed a module from scratch using TDD principles that approximates real-time market risk using live prices; and publishes out the result on a message queue. Helped retire a legacy system saving the firm ∼$250k.

Software Development Engineer — *July 2015 – Jan 2018*

- **Risk Management System**: Worked extensively on the core app used by traders for visualizing and hedging risk:
  1. Optimized the trades feed using LMax Disruptor, a low latency Java queue for upto 20% faster trades processing.
  2. Framework to validate critical live market data results which reduced manual testing effort by 90%.
  3. Mechanism to switch from a MongoDB replica set to standalone instance in the event of a data center failure.

Technical Skills

- **Languages**: Python, Java, Unix Shell Scripting, SQL, MATLAB
- **Frameworks**: PyTorch, TensorFlow, HuggingFace(Contributor), Scikit-Learn, Pandas, NumPy, NLTK, Swagger
- **Databases**: Sybase ASE, MongoDB, MySQL