

NEERAJ SHARMA

✉ neerajsharma@umass.edu ☎ (413) 695-7241 📍 Amherst, MA, US in linkedin.com/in/neeraj-sharma-86044a61 📄 github.com/neerajsharma9195
🌐 https://neerajsharma9195.github.io/

EDUCATION

MS in Computer Science (Concentration in Data Science)

University of Massachusetts Amherst

📅 Sep 2019 – May 2021 📍 Amherst, MA

- **Cumulative GPA: 3.86 / 4.0**
- **Courses taken:** Machine Learning, Systems for Data Science, Natural Language Processing, Algorithms for Data Science, Introduction to Simulation, Information Retrieval, Reinforcement Learning and Secure Distributed Systems.

Bachelor of Technology - Computer Science & Engineering

Indian Institute of Technology Mandi

📅 Aug 2012 – June 2016 📍 Mandi, India

- **Cumulative GPA: 9.03 / 10.0**
- **Relevant Courses:** Algorithms, Data Structures, Pattern Recognition.

WORK EXPERIENCE

LifeOmic

📅 May 2020 – Sep 2020 📍 Indianapolis, US (Remote- Amherst, MA)

Machine Learning Intern – Precision Health Cloud

- Improved the medical documents scanning and processing workflow by adding the **spell check and correction** step in OCR pipeline.
- Built a dictionary of 0.4 million words using [MIMIC-III](#), [PubMed](#) and [WordNet](#) dataset. Developed a statistical model based on [Damerau-Levenshtein distance](#) as baseline obtaining improvement on Word Error Rate by **14%**.
- Prepared a 5 million (noisy, corrected) sentence pairs dataset and trained a **Sentence-Piece tokenizer**.
- Trained an **attention based transformer** model with BPE based encoding. Model auto corrected **97%** word errors.
- Deployed statistical model as AWS lambda and deep learning model using AWS Fargate and integrated in OCR workflow.

MindTickle

📅 July 2016 – July 2019 📍 Pune, India

Research and Software Development Engineer II – ML Team

- Delivered an end to end solution for transcription, topic identification, calendar sync and text-search to the sales call analysis platform (CallAI).
- Classified sales manager feedbacks to subordinate's sales pitches by analyzing sentiment, sentence length, templates based identifiers as features in decision tree, obtaining **85%** accuracy.
- Annotated low confidence time-window in over **500** sales pitch audio and transcribed text using cues like silence gap, filler words, speech speed variations. Built a statistical model which identified low confidence windows with **75%** precision and **80%** recall.

Full Stack Developer – MindTickle Android App

- Developed a personalized notification service which increased the screen time of application by **30%**.
- Refactored the android application to improve scalability using MVP pattern and added new features like Courses, Mission and Coaching review modules.

SKILLS

- **Areas of Interest:** Machine Learning, Natural Language Processing and Distributed Systems.
- **Languages:** Python, JAVA, C++, SQL, Typescript, R, Matlab.
- **Machine Learning toolkit:** Deep Learning(PyTorch, TensorFlow), NumPy, SciPy, Stanford CoreNLP, NLTK, TextBlob.
- **Android Application Development:** Android Studio, Dagger 2, Retrofit, OkHttp Client.
- **Other tools & technologies:** Hadoop, Spark, Map Reduce, MySQL, PostgreSQL, Elasticsearch, Git, Docker, Kubernetes, Terraform, Helm charts, Google Speech-to-Text, AWS (S3, EC2, Dynamo, Redshift, Kafka, Lambda, Fargate, Transcribe).
- **Design:** Event and data-driven Systems, Microservice Architecture, MVC, REST.

PUBLICATIONS

Text classification using hierarchical sparse representation classifiers

Neeraj Sharma and A. D. Dileep and V. Thenkanidiyoor (ICMLA 2017)

- Built a **Hierarchical Sparse Representation Classifier (HSRC)** and explored it for text classification.
- Explored **weighted decomposition principal component analysis (WDPCA)** technique to highlight the middle principal components to construct the dictionary that emphasizes discrimination among confusing classes.
- HSRC classified **83.30%** of the documents correctly opposed to the **78.78%** (using SRC) with WDPCA on [20 Newsgroup corpus](#).

Text classification using combined sparse representation classifiers and support vector machines

Neeraj Sharma and Anshu Sharma and A. D. Dileep and V. Thenkanidiyoor (ISCB 2016)

- Explored frequency based kernels such as Histogram intersection kernel, χ^2 -kernel and Hellinger's kernel for text classification using SVMs.
- Combined the SRC and SVM classifiers based on voting scheme. Obtained **81.83%** accuracy using on [20 Newsgroup corpus](#).

RESEARCH

Complementary Count Min Sketch

Independent Study under Prof. Cameron Musco (Spring 2020)

- Proposed and analyzed a novel data structure called Complementary Count Min Sketch (CCMS) to support deletion in count min sketch.
- Compared the performance of CCMS with different sketches for feature selection in high dimensional data settings.
- Received comparable performance with Count Sketch under power and zipf's law distribution settings.

PROJECTS

Adversarial Recommendation Systems

Graduate Student Researcher (Amazon Alexa AI) (Spring 2021)

- We are working on addressing the cold start problem in recommendation systems using adversarial recommendation models utilizing reviews and ratings data of [Amazon review dataset](#).

Role of Attentive History Selection in Conversational Information Seeking

Academic Project (Information Retrieval) (Fall 2020)

- Experimented with soft-attention based conversational history selection approach in open retrieval setting on the [OR-QuAC](#) dataset.
- Compared the results with prepending based history selection [baseline](#) and performed ablation studies to assess the performance of individual components.

Programming Language Generation using Natural Language

Academic Project (Advanced NLP) (Spring 2020)

- Compared the performance of various neural network based architectures and identified the key factors for Python code generation such as beam search, multi headed attention on [CoNaLa dataset](#) and contextual [JulCe dataset](#).
- Obtained an improvement of **3.48 BLEU** score in transformer with back translation in comparison to standard transformer model on CoNaLa mined dataset.