Zeba Karishma

+1-(814)-852-9807 | zebakarishma@gmail.com | linkedin.com/in/zebakarishma | github.com/zebaKarishma

Education

The Pennsylvania State University — University Park, PA
M.S. in Computer Science and Engineering; CGPA: 3.8/4.0 — *Aug. 2019 – Aug. 2021*

Birla Institute of Technology,Mesra — Jharkhand, India
Bachelor of Engineering in Information Technology; CGPA: 8.5/10.0 — *Aug. 2012 – May 2016*

Experience

Yahoo! Inc. | *Location Platform Solutions Dept.* — Mountain View, CA
Software Development Engineer II — *Sept 2021 - Present*

- Currently enhancing an AI/ML-based ETL pipeline for **Yahoo Location Search**, processing 20M+ business listings for efficient ingestion, feature generation (text, numerical, geospatial), and deduplication/blending using an XGBoost classifier (86% precision, 78% recall) to improve data quality and retrieval accuracy.
- Enhanced a CNN-based category prediction framework for business listings by employing a custom loss function to learn category correlations, improving accuracy to 83% and effectively handling rare categories.
- Implemented a geo-spatial matching algorithm with 95% accuracy to match location geometries with the highest overlap using Apache Spark, Sedona, and Scala. Calculated spatial joins (intersect, contain, within) to determine parent/child/contain relationships and assigned permanent unique WOEIDs (Where On Earth Identifiers)

Penn State |The Intelligent Information Systems Research Laboratory — University Park, PA
Graduate Research Assistant — *Aug. 2019 - Aug 2021*
Advisor: Dr. C. Lee Giles and Co-advisor: Dr Jian Wu

- **Scientific Figures Extraction, Clustering and Classification (Master's Thesis):** Developed a deep-learning pipeline for extracting and classifying figures and tables from scientific literature using Vision Transformer (ViT) Model, which splits figures into fixed-size patches embedded with position embeddings and processed by a Transformer encoder, achieving 83% accuracy.[paper]
- **COVIDSeer:** Developed an AI-driven search engine for the CORD-19 dataset of research papers, featuring faceted search, keyphrase extraction, and similar paper recommendations using SciBERT embeddings. [paper]
- **Next-Gen-Citeseer:** Contributed to CiteSeer, a search engine for scientific papers, by implementing faceted search for new versions, including bucketing, clustering, aggregations, and filtering using Elasticsearch and Python [code]
- **Math Specialty Search Engine:** Indexed, retrieved, and ranked 1.1 million math text questions and formulas from the Mathematics Stack Exchange dataset using TF-IDF, BM25, and BERT re-ranking techniques.

Verizon Media |Big Data and Artificial Intelligence Dept. — Sunnyvale, California
Data Engineering Intern — *May 2020 – Aug. 2020*

- Developed and deployed a pipeline on a distributed cluster for analyzing customer's live location data to generate aggregation features on customer's location visits on a daily and monthly basis.
- Created a Datapack, a dictionary of unique POIs (such as airport, malls, etc.) and a pair of (lat, lng) and performed reverse geocoding on points and polygons to tag location data with their POIs.

Comviva — Bengaluru, India
Senior Software Engineer — *Dec 2018 - June 2019*

- **LEAP — USSD Application Development Platform (Research &Development)**
  - ∗ Developed an authentication server with user profile storage management, deployed across 19+ operator networks in Asia, the Middle East, and Africa. Created a generic macros module for USSD call flows, reducing parsing time by 30% and supporting complex payloads. Conducted onsite deployment for Du telecom operator in Dubai.
  - ∗ Engineered a session-based API Gateway for microservices, serving as a single entry point to enforce access control for applications, administrators, users, plugins, and reports. Implemented comprehensive workflow management for varying user permissions.

Comviva Bengaluru, India
Software Engineer *June 2016 - Dec 2018*

- **Real-time network analytics for better Customer Experience (Research & Development)**
  - ∗ Developed an algorithm to calculate network Quality of Experience (QoE) using parameters such as bandwidth, packet loss, round trip time, and cell handovers, and added an alerting module for monitoring network KPIs and sending alerts for KPI breaches in specific network cell IDs or regions.
  - ∗ Worked on an Android app for crowd-sourcing network probes, including developing an indoor/outdoor detection algorithm and sending collected data to the RACE server for customer experience analytics.

Technical Skills

Key Courses: Large Scale Machine Learning, Natural Language Processing, Data Mining, Information Retrieval, Computer Vision, Algorithm Analysis, Distributed Systems

Frameworks: TensorFlow, Keras, PyTorch, Pandas, Sklearn, Spark, Hadoop, Hive, Sedona, Docker, GIT, JIRA, AWS

Languages: Scala, Pyspark, Python, Java, SQL

Projects

Emoji Prediction: Crawled twitter data to create a dataset for emoji prediction for top 20 most used emojis on twitter and developed a prediction framework using ML and deep learning models. Finally, performed a comparative study on the effect of different features on model prediction.[paper]

Augmented Reality Viewer: A simple augmented reality viewer that displays artificial objects overlaid on images of real 3D scene.Its an offline implementation using COLMAP, RANSAC and 3D to 2D transfomations of world and camera coordinate systems.[code]

Object Tracking: Enhanced Siamese-RPN tracking with Mean Shift Algorithm and space scaling. Evaluated the appearance model at multiple resolutions by performing an exhaustive scale search to incorporate scale estimation in a tracking framework. [code]

Spectral Graph Convolutions in GCN: A generalization of CNNs from low-dimensional regular grids to signals defined on more general high dimensional domains like graphs.Performed a mathematical study to show how fast convolutions are achieved by first-order approximation of spectral graph convolutions can therefore be used to build a neural network model.[paper]

Anomaly Detection in Bipolar Disorder: Developed a prediction framework that will identify behavioral anomalies and early-warning signs in Bipolar Disorder using continuous streams of online behavioral data.

Toxic Comment Classification using Wikipedia's talk page edits: Implemented an ensemble of deep learning models to build a multi-headed model capable of detecting multiple types of toxicity to help improve online conversation.[code]

Publications

For full list of publications, please visit: google scholar, semantic scholar.

Z. Karishma. **Scientific Document Figure Extraction, Clustering and Classification**, 2021. URL https://etda.libraries.psu.edu/catalog/19941zbk5052.

Z. Karishma, S. Rohatgi, K. S. Puranik, J. Wu, and C. L. Giles. **ACL-Fig: A Dataset for Scientific Figure Classification**. In *Proceedings of the Workshop on Scientific Document Understanding (SDU 2023) co-located with 37th AAAI Conference on Artificial Inteligence (AAAI 2023)*, 2023. URL https://doi.org/10.48550/arXiv.2301.12293.

S. Rohatgi, Z. Karishma, J. Chhay, S. R. R. Keesara, J. Wu, C. Caragea, and C. L. Giles. **COVIDSeer: Extending the CORD-19 Dataset**. In *Proceedings of the ACM Symposium on Document Engineering 2020*, DocEng '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380003. doi: 10.1145/3395027.3419597. URL https://doi.org/10.1145/3395027.3419597.

P. Venkit, Z. Karishma, C.-Y. Hsu, R. Katiki, K. Huang, S. Wilson, and P. Dudas. **A 'Sourceful' Twist: Emoji Prediction Based on Sentiment, Hashtags and Application Source**, 2021. URL https://doi.org/10.48550/arXiv.2103.07833.

Honors and Rewards

Best Paper Award among 543 submissions *at Yahoo! Techpulse* Mountain View,CA
For paper "Identifying Geoinformatics Data and Discovering Spatial Relationship at Scale" *April 2023*

Graduate Research Assistantship State College, PA
Partly supported By National Science Foundation grant under award 1823288. *June 2020*

Best Team Award at Comviva Bengaluru, India
Awarded to a team of 5 for designing, developing and successfully deploying LEAP in 19+ countries *Mar 2019*

Discretionary Award at Comviva Bengaluru, India
For handling product deployment for operator Telefonica, Spain single-handedly as fresher. *June 2017*

cit Venkit et al. [2021] Rohatgi et al. [2020] Karishma et al. [2023] Karishma [2021]