

# Senthil Kumar

## Applied ML Data Scientist

Competent in the use of ML and DL for NLP tasks || Proficient in building clean, modular and **dockerized Python applications** || Of the total 11+ years of experience, 9.5 years of individual contribution and 1.5 years of people management || [Detailed Profile](#)



senthilkumar.m1901@gmail.com

+91 984-186-9609/ +91 944-512-3824

linkedin.com/in/senthilkumarm1901

github.com/senthilkumarm1901

## WORK EXPERIENCE

### Deputy Manager

#### Ford Analytics Division

05/2018 - Present

Chennai, India

##### Roles

- **Senior Analyst [1.5 Years] | Deputy Manager [2+ Years]**
- **A data science developer** who employs state-of-the-art ML and DL techniques for NLP Applications
- -- for analytics teams such as AI Advancement Center, Customer Experience and Manufacturing Operations
- Experienced in end-to-end ML application development
- -- from data acquisition, cleaning, labeling and preprocessing,
- -- to model development, deployment and maintenance
- **Python/NLP Trainer | Technical Interviewer** of NLP candidates across analytics teams

### Assistant Manager

#### LatentView Analytics

04/2014 - 04/2018

Chennai, India

##### Roles

- **Individual contributor [2.5 years] | People Manager [1.5 years] | Social Media Analysis | NLP Projects**
- LinkedIn Recommendation: "... extraordinary dedication contributed significantly to growing our analytic practice..." - **F100 Tech Client stakeholder**
- LinkedIn Recommendation: "...Sincere, driven, articulate and utterly committed ..." - **Skip-level Reporting Manager at LatentView**

### Senior Consultant

#### Capgemini

01/2014 - 03/2014

Bangalore, India

### Lead Analyst

#### Beroe Inc - A Procurement Intelligence Firm

07/2010 - 12/2013

Chennai, India

##### Achievements/Tasks

- LinkedIn Recommendation: "... well organized, innovative ... and always ready to go the extra mile to support the client ..." - **Client Engagement Manager**

## EDUCATION

B.E - Electronics & Communication - 8.6 CGPA

Madras Institute of Technology , 2006 - 2010

Deep Learning Specialization

DeepLearning.AI-Coursera (5 courses), Dec 2018 - May 2019

Applied ML and Applied Text Mining

University of Michigan - Coursera, Dec 2017 - Jan 2018

2 Refresher Courses on ML Fundamentals for Structured Data

Kaggle Learn, Feb 2021

12th Grade - 95% | 10th Grade - 92%

State Topper in Physical Science paper of 2006 TN Engineering Entrance Exam

GCP Big Data & ML Fundamentals

Google - Coursera, Apr 2021

Refresher Course on SQL (BigQuery) Fundamentals

Kaggle Learn, Feb 2021

Probability and Statistics Fundamentals

LinkedIn Learning, Dec 2021

## TECHNICAL SKILLS

### Languages

Python, Markdown, Linux Shell (basics), SQL (basics)

### Tools

Git, WSL, Docker, Kubernetes, Poetry (Python env), Conda, PyCharm

---

## TECHNICAL SKILLS

---

**Python Libraries  
(extensive usage)**

Pandas, SpaCy, Re (Regular Expressions),  
Transformers, Sklearn, PyTorch

**Python Libraries  
(working  
knowledge)**

PySpark, FastAPI (REST API), Streamlit (UI),  
Altair (viz)

---

## PROMOTIONS AND AWARDS

---

**Ford: Promoted from Senior Analyst to Deputy Manager**

*In Nov'2019, after 1.5 years of joining Ford*

**Ford Asia-Pacific Recognition Award**

*won in May 2019 for successful spearheading of a project*

**LatentView: Promoted from Senior Analyst to Assistant Manager**

*In Oct'16, after 2.5 years of joining LatentView*

**LatentView Analytics - Encore Award**

*won for company-wide best performance for the Jul-Sep 2016 quarter*

**Beroe: Promoted twice in my first company**

*During my 3.5 year stint in Beroe*

**Beroe - Knowledge Contributor Awards**

*Won twice for company-wide best performance in Q1 and Q2 calendar year 2013*

---

## SAMPLE KEY PROJECTS

---

**(1) BERT Fine-tuned Aspect-based Sentiment Analysis (more details in link -->) [↗](#)**

- **Goal:** To Build a reusable Sequence Classification ML Pipeline
- -- which converts customer comments into trackable Aspect and Sentiment pairs
- The development environment, replicable via docker for model training and inference,
- -- (1) is used for building 30+ different text classification models
- -- (2) is used by analysts with limited knowledge in DL
- The pipeline helped in
- -- (1) less annotation for Training (compared to a traditional ML algorithm) by intelligent use of DL+ML models
- -- (2) achieving easily an F1 score of 85%+ for all models with tough 25+ classes and with just 2K-4K annotated data
- **Libraries:** Python, PySpark, SpaCy, PyTorch; **Tools:** GPUs, Shell scripting, Docker, PyCharm and GitHub

**(2) Personally Identifiable Information (PII) Detection using NER (more details in link -->) [↗](#)**

- **Goal:** To anonymize PII in text data
- -- (1) by building a Named Entity Recognition (NER) system which employed both RoBERTa Transformer model and Rules-based logic
- -- (2) by replacing the PII words with appropriate generic text
- -- (3) that can result in less restricted use of the data
- Bootstrapped the training data using Spacy rules (thus easing the annotation process by not starting labeling from scratch)
- Achieved an F1 score of 89% for detecting the PII entities
- Deployed an asynchronous\* Inference REST API (using FastAPI and K8s) that can be plugged into multiple applications
- **Libraries:** Python, SpaCy, Transformers, PyTorch, Celery/Redis\*, FastAPI; **Tools:** GPUs, Docker, Kubernetes, PyCharm, and GitHub

**(3) Unsupervised NLP Semantic Search Pipeline**

- **Goal:** To connect two automotive domain specific text data sources,
- -- which has technician comments about issues before the launch of a vehicle,
- -- by assigning semantically matching common part categories to every issue in both data sources
- Built a pipeline that ensembles results of 3 pairs of Retriever-Reader models wherein
- -- the Retriever narrows down the search space and
- -- the Reader zeroes in on the right results
- Built a simple **Streamlit UI** interface for a business user to try the approach and
- Created a **CLI app** for a domain analyst to experiment with different hyperparameters in parallel across different GPUs
- **Libraries:** Python, Sentence Transformers, PyTorch, SpaCy, Streamlit; **Tools:** GPUs, Docker, PyCharm, and GitHub

**(4) Text Data Clustering Pipeline (more details in this link -->) [↗](#)**

- **Goal:** To build reusable text data clustering pipeline
- -- (1) with simpler Python APIs for non-NLP analysts,
- -- (2) for deriving actionable insights from unlabeled text corpus
- The clustering pipeline provided options for both Traditional Topic Modeling and DL-Embedding powered Hard Clustering
- Incorporated the models into an easy-to-use Streamlit UI deployed via K8s
- **Libraries:** Python, Sentence Transformers, Transformers, Sklearn, Seeded LDA, pyLDAvis, Streamlit; **Tools:** GPUs, Docker, Kubernetes PyCharm, and GitHub

**(5) Miscellaneous Adhoc Efforts**

- **Social Media Analysis:** Analyzed the latent preferences expressed by consumers owning different vehicle models in Reddit threads for aiding in targeted marketing
- Built a **multi-GPU inferencing pipeline** for enabling parallel prediction of a Neural Machine Translation model (developed by a different team)
- **Benchmarking of Speech2Text Models:**
- -- Explored Speech2Text models such as DeepSpeech (open source), Microsoft Speech2Text API (paid) and Google Speech2Text API (paid)
- -- Compared their Word Error Rate performance for open source LibriSpeech and company-internal speech datasets
- **Migrated docker environment** and codebase of projects from 2019 to be compatible with Cuda 11 as company upgraded its on-premise GPU infrastructure
- Regularly aid fellow team members in the Kubernetes deployment of their applications