

SOHAM RAY

(917)753-6259

sr2259@cornell.edu

[linkedin.com/sohamray19](https://www.linkedin.com/sohamray19)

github.com/sohamray19

EDUCATION

Master's in Computer Science, Cornell University

Aug 2019 - May 2020

- **GPA: 3.86/4** | Courses: Large Scale Machine Learning, Natural Language Processing, AI, Computational Linguistics

Bachelor's in Information Technology, University of Pune

Jul 2015 - Jun 2019

- **GPA: 8.51/10** | Courses: Analysis of Algorithms, Data Structures, Cloud Computing, Databases, Computer Networks
-

TECHNICAL SKILLS

Experience in building web apps/ML models in Java/Python with strong competencies in data structures, algorithms, object oriented design, large scale data processing, and scrum; supplemented with expertise in machine learning and nlp.

- Languages: **Java, Python**, C++, Scala, Solidity, SQL, HTML, CSS, Javascript
 - Frameworks: **Data ETL** (Apache Kafka, Spark, Flume, Avro, Hive, AWS), **Web Dev** (Maven, Git, Java EE, Flask, RESTful services, Eclipse), **Machine Learning** (Tensorflow, Keras, NLP frameworks, Scikit), **Blockchain** (Hyperledger, Ethereum)
-

EXPERIENCE

Software Engineering Intern, Cybage, Pune, India

Feb 2019 - May 2019

- Developed a software to analyze sentiment and abstracted the product domain so it can be used for any feedback analysis task
- Improved classification accuracy by **11%** on inplace system, and scored an accuracy of **80%** on data from un-trained domains

Project Intern, ASquared IoT, Pune, India

July 2018 - May 2019

- Developed a web application in python to classify welding sound files based on quality with a projected accuracy of **71%**
- Constructed denoising models using statistical and ML methods to improve projected classification accuracy to **84%**

Software Engineering Intern, Exadatam, Pune, India

June 2018 - Dec 2018

- Developed a production-ready real-time streaming ETL data pipeline, with pluggable ML model for sentiment analysis
- Designed data visualization dashboards illustrating various Key Performance Indicators (KPIs) of a Fortune 100 Company

Software Engineering Intern, Exadatam, Pune, India

Dec 2017 - Mar 2018

- Developed two prototypes of a distributed blockchain application in supply chain and logistics management
 - Comparatively analyzed prototypes on the Ethereum blockchain and Hyperledger Fabric to choose the better framework
-

PROJECTS

Master's Project in Data-driven, Hyper-local Air Quality Modeling, Cornell University

- Built data engineering pipelines and ML models to predict hotspots and observe patterns (in air pollution) in global metropolitan areas (London and Oakland) to help curb rising air pollution levels worldwide. RMSE: **0.06**

Insurance Policy Database Management System, University of Pune

- Constructed a web application with role-based access control, interactive payment reminders, payback estimate on insurance maturity, and usage analysis reports for customers/agents to ensure one-stop solution

Deceptive Opinion Spam Classification, Cornell University

- Constructed an opinion spam classifier that is able to weed out fake reviews written by real people. Explored usage of language models and naive bayes classifier using different linguistic and preprocessing techniques. Accuracy: **91%**

Metaphor Detection with Sequence Labeling Models, Cornell University

- Implemented part of speech tagging and metaphor detection on a given text corpus. Learned feature engineering, hidden markov models and feedforward neural networks to get the observation probability matrix. F1 score: **66%** Accuracy: **85%**

Story Completion Prediction (Cloze Test), Cornell University

- Worked with feature selection and engineering, classification models and pretrained language models (OpenAI GPT-2) to correctly predict the correct ending of a 4 story sentence given 2 endings. Accuracy: **87%**

Classification of Donald Trump's Tweets, Cornell University, Ithaca, NY

- Constructed an ML model to evaluate the hypothesis that Trump's tweets from his iphone are written by him and those from his android are written by his team, by classifying them using NLP techniques. Accuracy: **86%**
-

CERTIFICATIONS

- Coursera: DeepLearning.ai Specialization
- Udemy: Data Science for Professionals