**Sofia Dutta**

**Data Scientist | Experienced Software Developer**

443-554-4170 • sofiad1@umbc.edu • linkedin.com/in/sofiadutta • github.com/sofiadutta

**Education**  
University of Maryland, Baltimore County (UMBC), Baltimore, MD      Spring 2019 – Fall 2020

Master’s Professional Studies, Data Science, GPA: 4.0

West Bengal University of Technology, Kolkata, India Fall 2006 – Spring 2010

Bachelor of Technology, Computer Science, GPA: 3.5

**Technical Skills**

Programming Languages: Python, SQL, PL/SQL, T-SQL, Java

Development Tools: Docker, Jupyter Notebook, Google Colab, PL/SQL Developer, Git

Enterprise Tools: Google Cloud Platform, Amazon Web Services S3, Oracle Applications

Backend Tools: Oracle (9i, 10g, 11g), Microsoft SQL Server, MongoDB, JSON

Data Tools: PyTorch 1.4, Sci-kit Learn, Apache Spark 2.4, MLlib, Keras 2.2, Tensorflow 1.15,

Operating Systems: Mac OSX, Windows, Linux (Ubuntu, RHEL)

**Graduate studies projects**

Big Data Twitter Stream Sentiment Analysis Fall 2019

* Learned to use Twitter data APIs. Collected tweets, then cleaned and pre-processed using Python’s libraries.
* Used Apache Spark streaming API to collate data and applied Map-Reduce operations to track trending cryptocurrency topics.
* Visualized sentiment movements on trending cryptocurrency topics to find out if “humans on Twitter” are feeling positive about Bitcoin’s future or are they feeling negative.
* Visualized geographic distribution of tweets for trending cryptocurrency topics
* Created my own sentiment classifier by training on popular 1.6 million tweet data set
* Compared my classifier and well-known social media classifier. Found we agreed 7/10 times.

Sentiment Analysis on user review datasets from Amazon and IMDb Spring 2019

* Compared performance of traditional machine learning algorithms like support vector machines, logistic regression, versus neural networks created using Keras CNN, Keras Bidirectional LSTM to empirically prove neural networks are better at sentiment classification

Data characterization projects using Python Sci-Kit Learn         Spring 2019

* Analyzed Baltimore City Employee Salary data to prove there is no income inequality in Baltimore City Government
* Studied New York City Film Permits data to figure out top filming locations for popular movies
* Combined two different datasets from the New York City Fire Department and showed that it is possible to use data analysis techniques to find high impact incidents

**Publication**

Sofia Dutta et. al., “Context Sensitive Access Control in Smart Home Environments”, InProceedings, *6th IEEE International Conference on Big Data Security on Cloud (BigDataSecurity 2020)*, May 26, 2020, Baltimore, MD, USA.

**Work Experience**

Ebiquity Research Group, UMBC, Student Researcher, Baltimore, MD     Sep 2019 – May 2020

* Working on research in the area of Semantic Web, Context-based Access Control and Smart Home Automation
* Submitted a paper to IEEE Big Data Security 2020 conference.

Tata Consultancy Services (TCS), Kolkata, India                   Nov 2010 – Feb 2018

* Led a team of developers in preparing PL/SQL stored procedures.
* Designed, developed and tested API interfaces for PL/SQL stored procedures.
* Prepared functional specification documents.
* Performed requirement and change based regression analysis.
* Prepared test plans and performed system integration testing and user-acceptance testing.
* Worked on Oracle Fusion HCM (Core HR) functionalities for clients
* Wrote migration scripts for customer data migration projects.
* Completed client data migration from legacy Oracle Apps (11i) to Oracle ERP Suite (R12).
* Managed production environment deployments.

Relevant Coursework  
Data 601: Introduction to Data Science Spring 2019

* Performed data analysis projects using supervised and unsupervised machine learning packages.
* Worked on data collection, storage, transformation, cleaning, analysis, and visualization.

Data 602: Introduction to Data Analysis and Machine Learning Spring 2019

* Worked on practical machine learning and data analysis problems.
* Worked on end-to-end processing pipeline for extracting and identifying useful features that best represent data, applying machine algorithms, and evaluating their performance for modeling data.
* Learned machine learning APIs like Sci-kit Learn, Keras, Tensorflow.
* Learned machine learning algorithms like decision trees, logistic regression, support vector machines, convolutional neural networks, recurrent neural networks, bidirectional LSTM.

Data 603: Platforms for Big Data Processing         Fall 2019

* Using Apache Spark performed Map-Reduce operations on streaming data
* Learned Big Data technologies like PySpark, Spark SQL, MLlib, Spark Streaming, Hive, Hadoop
* Worked on practical projects with large datasets
* Used NoSQL storage system (MongoDB) to manage large datasets collected from Twitter Data APIs

Awards and recognitions

Awarded certificate: “Recognition of exceptional leadership and teamwork skills” at school 2019

Awarded TCS Gems “Champions of ILP” for contributions in training new employees 2017

Awarded certificate: “Most likely to Slay a Dragon” for efforts in Finance Data Migration 2015