**AQSA SHEIKH**

+1 (862)-256-9894 | aqsa2777@gmail.com| [linkedin.com/in/aqsasheikh27/](https://www.linkedin.com/in/aqsasheikh27/) | [github.com/aqsa27](https://github.com/aqsa27/)

# PROFESSIONAL EXPERIENCE



**Education Testing Services, Princeton, NJ, USA June 2022- March 2023**

**Data Engineer/ AI Engineer**

***ETL:***

* Developed **ETL** processes for ingesting data from diverse data sources into the **data warehouse** using **Python** and **AWS Glue**, resulting in improved data accessibility and analysis capabilities
* Loaded processed data into **S3** and **Redshift** clusters, ensuring data is readily available for analysis and decision-making
* Created **Airflow DAGs** to ensure data integrity and availability for training and scoring pipelines, while monitoring and debugging pipeline issues using web UI, logs, and metrics to ensure data quality and reliability.
* Utilized **SQL** and **PySpark** in **AWS Glue** to develop complex **data transformations**, enabling **data cleansing**, **normalization**, and **aggregation** from various sources
* Created interactive **dashboards** and reports in **Tableau** for business stakeholders, allowing for quick and effective decision making based on real-time data analysis
* Designed and implemented **CI/CD pipelines** for data engineering and analytics projects using **Bamboo**, **Git**, and **Docker**, resulting in a 50% reduction in deployment time and improved collaboration between teams.
* Independently designed and developed data collection scripts using **Python** and **Bash scripting**, resulting in increased efficiency and accuracy of data collection. Utilized advanced data cleaning techniques and tools to aggregate and clean data, ensuring high quality and consistent data for analysis
* Ensured data governance and compliance by **documenting and maintaining technical documentation**, **data dictionaries**, and **data lineage diagrams** for data systems

***AI/ML:***

* **Contributed to engine and algorithm development** for English speaking, writing, and reading assessment models, resulting in improved accuracy and effectiveness of the models.
* Lead end-to-end **machine learning (ML) lifecycle** using **MLflow**, encompassing problem framing, large-scale data collection, exploratory data analysis, model development, and performance measurement, to drive effective solutions
* Trained and deployed machine learning models to generate predictions and track performance metrics, utilizing open-source tools such as **Rater Scoring Modeling Tool (RSMTool)** and **SKLL** for evaluating automated scoring models and producing customizable **HTML statistical reports**
* Collaborated with analysts and business stakeholders to deliver data-driven solutions for reporting and analytics
* Developed and executed automated unit tests using the **pytest** and **unittest** frameworks in **Python** to validate the accuracy and reliability of machine learning models

***Analysis and Operations:***

* Effectively **communicate** progress and interpretation of experimental results to technical and business stakeholders, ensuring alignment and understanding
* Contribute to an **agile development** environment by participating in **user story creation**, **sprint planning**, and **review and retrospective meetings**

**ClearOps Inc, New York City, NY, USA Oct 2019 – May 2022**

**Computer Programmer (NLP/Machine Learning)**

## Patent Approved for DAWN (Automated A.I based Question Answering System)

***ETL***:

* Using **AWS Lambda,** implemented end-to-end pipeline for extracting questionnaire data, **cleaning and pre-processing CSV and Excel** file into format required by the NLP Model using **Python** and loading processed data into **S3 buckets**
* Worked on **data cleaning** and ensured data quality, consistency, integrity using **Pandas, NumPy, Python**
* Experience writing **Production level code** and applying state-of-the-art NLP models to find solutions to business problems
* Implement, evaluate and deploy NLP Automation Pipeline using Amazon Web Services like **S3, EC2, Dynamo DB and Lambda**
* Building **API’s** using **Flask** to deploy the different NLP model as endpoints to be consumed by the Web application

## NLP:

* Researching, Studying and building Deep Learning Model in Python that will analyze data in Question Answering format using Deep Learning Framework like **BERT, Transformers, GPT-3, GPT-2, Albert**
* Build multiple Deep Learning Model for NLP task such as **Question classification, Reading Comprehension, knowledge Graph based Question Answering System, Semantic Analysis, Text Analytics**
* Built Promise module using **NER** technology that extracted future promises done by companies while answering Security Questionnaire
* Build and integrated two end-to-end Automated A.I powered Question Answering System using Deep Learning frameworks like **BERT and Transformers** into the Web Application with **Feedback Optimization**
* Extensively worked using machine learning libraries such as **Spacy, Stanza, NLTK, Seaborn, SciKit-Learn, SciPy** for machine learning

# Developed NLP models for Named Entity Recognizer, Grammar & Spell Check, Entity Extraction, Sentiment Analysis, Question Generation, Text Summarization

* Constructed a language model to understand the semantics of the data and apply various NLP task on the data

## Analysis and Leadership:

* Actively gathered requirements from Clients to turn Business problem into technical models
* Responsible for **supervising and mentoring** Interns on the team
* Collaborated with team members for **task planning, task distribution** and for integration of models
* Understand the problem statement, client requirements and design and build complex solutions using Programing languages and state-of-the-art NLP models
* Developed user stories in **JIRA** according to new requirement and potentials change requests

## Operations:

* Setting up **alerts** and notification for pipeline failures utilizing **CloudWatch** logs
* Responsible for initial client interaction for production issues, triaging them and providing solution, handling change Requests, facilitating analysis and optimizing the models according to the feedback

# V group Inc, Princeton, NJ, USA

**NLP Intern Sep 2019 – Oct 2019**

* Build **chatbot** by Integrating **Microsoft Azure** cognitive service **LUIS** with **QandA maker**
* Developed product solution using NLP libraries like **Gensim, NLTK, Vader, SpaCy**

**PROJECTS**



* **Daily Birth Forecasting Time Series |Python, NLTK, Scikit-learn, Flask:** [Disaster\_message\_classifier](https://github.com/aqsa27/Disaster_message_classifier)

A multilabel classification model to predict the categories of a disaster message. Includes an ETL pipeline for data processing, a ML pipeline to train the model, and a web app, with visualizations, where the model can be used to classify messages.

* **Twitter Hate Tweet Classification** |***NLP, Python****:* [Text\_Classification\_NLP](https://github.com/aqsa27/Text_Classification_NLP)

Implemented a model for analyzing negative tweets utilizing NLP text classification and Naïve based model

* **Files Manager** | ***AWS (S3, Lambda, DynamoDB, API Gateway), Dash, Python****:* [AWS\_Visualization\_Manager](https://github.com/aqsa27/AWS_Visualization_Manager) Pipeline for storing files along with meta-data, files can be searched and visualize using metadata
* **Optimizing Apache Lucene Search Engine**| ***Java****:* [Search\_Engine\_Lucene](https://github.com/aqsa27/Search_Engine_Lucene)

Search engine with 1.65% high relevancy score implemented using Lucene, Spell Checker and snippet

* **Spotify-artists-analysis | R:** [spotify-artists-analysis](https://github.com/aqsa27/spotify-artists-analysis)

The purpose of this project is to analyze how different or how similar is the music that different artist on Spotify produce.

# EDUCATION

|  |  |
| --- | --- |
| **Master’s in information systems** |  |
| **New Jersey Institute of Technology**, Newark, NJ, USA | **Dec 2018** |
| **Bachelor’s in biotechnology** |  |
| **Mumbai University**, Mumbai, India | **May 2016** |

# SKILLS

* **Programming and Scripting Languages:** Python, SQL, R, PySpark and Bash
* **Cloud Platforms:** Amazon Web Services, Azure and GCP
* **ETL Tools:** Apache Spark, Apache Airflow, AWS Glue, Talend, Google Dataflow
* **Relational databases:** MySQL, PostgreSQL, Oracle
* **NoSQL databases:** Cassandra, DynamoDB
* **Stream Processing:** Apache Kafka, Apache Spark Streaming, AWS Kinesis
* **Machine Learning Frameworks:** TensorFlow, PyTorch, HuggingFace, Scikit-Learn
* **DevOps and automation:** Docker, Kubernetes
* **Data Visualization Tools:** Tableau, Power BI, Matplotlib, Seaborn, ggplot2, QlikView, KNIME
* **Data Warehousing:** Amazon Redshift, Google BigQuery, Snowflake
* **Natural Language Generation (NLG) Tools:** GPT-3, BERT, ChatGPT-4
* **Version Control and CI/CD Tool:** Git, Github, BitBucket, Bamboo, JIRA, Gitlab
* **Web Frameworks:** Flask, Django, HTML, CSS, Dash by Plotly

**CERTIFICATIONS**



* Machine Learning Data Lifecycle in Production
* Deep Learning: Advanced NLP and RNNs
* Microsoft: DEV288x: Natural Language Processing (NLP)
* DAT207x: Analyzing and Visualizing Data with Power BI
* DE with **Google Cloud Platform** including Data Analysis with Google Big Query and Cloud Dataflow
* Big Data and Machine Learning Fundamentals with Tensor Flow on Google Cloud Platform