Hassan Abdelfattah Obaia

Junior Data Scientist – AI and NLP Practitioner

hassanobaya@gmail.com | Mobile (EG): (+20) 1550123886 | Mobile (IT): (+39) 3204079413 | Milan, Italy
GitHub: hassan-obaya | LinkedIn: hassan-obaya
Nationality: Egyptian | **Gender:** Male | **Military Status:** Exempted

OBJECTIVE

Aspiring Machine Learning Engineer and Data Scientist with hands-on experience in developing end-to-end AI solutions through real-world projects, industry internships, and mentoring roles. Skilled in supervised and unsupervised learning, deep learning, and statistical analysis. Proven ability to build and evaluate models using Python, scikit-learn, PyTorch, and cloud tools. Seeking a challenging opportunity to contribute to impactful AI projects, combining technical excellence with a strong foundation in data-centric problem solving.

TECHNICAL SKILLS

Programming Languages: Python, SQL, Bash
Machine Learning: Supervised Learning (Regression, Classification), Unsupervised Learning (Clustering, Anomaly Detection), Model Evaluation (F1, AUC, RMSE), Model Tuning (GridSearchCV, RandomizedSearchCV), Feature Engineering
Deep Learning: LSTM for Time Series Forecasting, BERT for NLP (Hugging Face Transformers), Keras
Natural Language Processing: Text Classification, Sentiment Analysis, Tokenization, TF-IDF, Bag-of-Words, Transformers, SQuAD QA, Retrieval-Augmented Generation (RAG), Sentence Transformers, FAISS
Statistical Analysis: Hypothesis Testing (Mann-Whitney U, A/B Testing), Descriptive Statistics, Correlation Analysis, ARMA and GARCH Models
Data Analysis & Visualization: Pandas, NumPy, Matplotlib, Seaborn, Plotly, Dash, EDA, Heatmaps, Choropleth Maps, Time Series Analysis
Deployment & Tools: Flask (Web Deployment), Streamlit (UI for ML Apps), Jupyter Notebook, Git, GitHub
Data Sources: CSV, SQL Databases, NoSQL, REST APIs
Libraries & Frameworks: Scikit-learn, XGBoost, Statsmodels, Keras, Hugging Face Transformers, FAISS
Other Tools: Google Colab, Kaggle Notebooks, VS Code, Linux CLI

EDUCATION

Bachelor of Science in Artificial Intelligence — Kafr El-Sheikh University, Egypt
September 2021 – July 2025

- Relevant Coursework: Machine Learning, Deep Learning, Data Science, Computer Vision, NLP

Applied Data Science Lab, WorldQuant University — Fully online, Credly-certified
16-week project-based credential
Oct 2024 – Jan 2025

- Completed 8 end-to-end data science projects covering data extraction, cleaning, analysis, modeling, and deployment.
- Worked with diverse data sources (CSV, SQL, NoSQL, APIs) using Python and industry-standard libraries.
- Applied models including Linear Regression, Random Forest, K-Means, ARMA, and GARCH on real-world problems (e.g., housing prices in Mexico, air quality in Kenya, earthquake damage classification).
- Performed A/B testing, statistical analysis, and data visualization using tools like Plotly Dash, Seaborn, and Matplotlib.

TECHNICAL EXPERIENCE

Applied NLP Engineer - Intern — Elevvo Pathways, Remote
July 2025 – Present

- Developed end-to-end NLP pipelines in Python, including data cleaning, preprocessing, and feature engineering.
- Trained and evaluated models for sentiment analysis, fake-news detection, named entity recognition, and topic modeling.
- Fine-tuned transformer models for tasks such as question answering and summarization.
- Built semantic resume-screening tools to match candidate profiles with job descriptions.
- Tools: Scikit-learn, NLTK, spaCy, Transformers, Hugging Face, Pandas.

Machine Learning Engineer - Intern — Cellula Technologies, Remote
February 2025 – March 2025

- Built, trained, and deployed machine learning models for data-centric applications, enhancing predictive performance.
- Contributed to AI research and implemented scalable model pipelines for production environments.
- Collaborated with teams to deploy models using scikit-learn.

Machine Learning Scientist Intern NeuronetiX, Remote
Aug 2024 – Oct 2024

- Applied supervised learning techniques (Random Forest, Ridge Regression, XGBoost) to solve classification and regression tasks on real-world datasets.
- Implemented unsupervised learning methods such as K-Means and PCA for clustering and dimensionality reduction in exploratory settings.
- Conducted data cleaning, feature engineering, and statistical analysis to prepare datasets for machine learning workflows.
- Documented experiments thoroughly and supported the transformation of research models into deployable prototypes.

Machine Learning Scientist - Intern Shai AI, Remote
Aug 2024 – Oct 2024

- Gained practical experience with both supervised and unsupervised learning models using real-world datasets.
- Implemented core algorithms including decision trees, k-nearest neighbors (KNN), and k-means clustering using Python and scikit-learn.
- Evaluated model performance using accuracy, silhouette score, and confusion matrices.
- Improved understanding of model selection, feature importance, and hyperparameter tuning in classical ML workflows.

PROJECT HIGHLIGHTS

End-to-End ML Deployment – Hotel Booking Cancellation Prediction GitHub
Technologies: Flask, Python, Scikit-learn, Random Forest, Pandas, Docker, Git

- Engineered a full ML pipeline: data extraction, cleaning, imputation, one-hot encoding, normalization, and feature selection.
- Developed and Dockerized a Flask web application with RESTful API endpoints for batch uploads and real-time predictions.
- Optimized model performance via GridSearchCV and 5-fold cross-validation, achieving 89

LLM-Based QA System with PDF and TXT Support GitHub
Technologies: LangChain, Hugging Face Transformers, FAISS, Sentence Transformers, Streamlit

- Architected a Retrieval-Augmented Generation (RAG) pipeline with FAISS vector store for semantic search over large document corpora.
- Built PDF/TXT ingestion and preprocessing workflows, improving top-k retrieval relevance by 15
- Deployed an interactive Streamlit interface integrating the Mistral-7B model for contextual answer generation.

Question Answering with Transformers – SQuAD v1 GitHub
Transformers · Hugging Face · BERT · PyTorch · NLP

- Fine-tuned DistilBERT on the SQuAD v1.1 dataset for extractive QA using Hugging Face Transformers.
- Deployed a QA pipeline with support for both sample and custom questions using GPU acceleration.
- Evaluated with Exact Match and F1 metrics; handled model/tokenizer integration and dataset loading via `datasets`.

News Category Classification – AG News GitHub
Technologies: Scikit-learn, TensorFlow, TF-IDF, Neural Networks

- Developed multiclass classifiers using TF-IDF features with Logistic Regression and a custom TensorFlow DNN.
- Conducted stratified k-fold cross-validation and grid search, achieving 93
- Analyzed confusion matrices and ROC curves to address class imbalance and refine feature selection.

Stock Price Forecasting – Apple (2015–2023) GitHub
LSTM · Time Series · Deep Learning · yfinance · MAE/RMSE

- Developed a stacked LSTM model to forecast Apple's daily closing price using historical OHLCV data and technical indicators (e.g., MA, RSI).
- Automated data collection via `yfinance`, performed feature engineering and sequence generation for supervised learning.
- Achieved RMSE of 6.76 USD on test set; visualized predictions across train/val/test splits with Matplotlib.

Decades of Hunger: A Global Look at Famine Deaths GitHub
Technologies: Pandas, Scikit-learn, Isolation Forest, Plotly, Seaborn, Folium

- Performed exploratory data analysis on historical famine datasets with correlation matrices and hypothesis testing.
- Applied Isolation Forest for anomaly detection of extreme famine events, correlating outliers with GDP per capita.
- Designed interactive choropleth maps and time-series dashboards in Plotly and Folium to visualize spatial-temporal trends.

CERTIfiCATIONS

- Llama Fundamentals – DataCamp *Jun 2025*
- Machine Learning Scientist – DataCamp *May 2025*
- Nature Language Processing in Python – DataCamp *May 2025*
- Data Scientist in Python – DataCamp *April 2025*
- Associate Data Scientist – DataCamp *April 2025*
- Machine Learning Certification – Shai AI *Oct 2024*
- Power BI Fundamentals – DataCamp *May 2024*
- Data Analyst with Python – DataCamp *April 2025*

ACTIVITIES

Hackathon Participant NeuronetiX Hackathon, Cairo
Jan 2025

- Collaborated with a cross-functional team to develop machine learning models for predicting telecom customer churn.
- Conducted EDA, feature engineering, and applied classification algorithms (e.g., Logistic Regression, Random Forest, XGBoost).
- Achieved strong F1-score and ROC-AUC on imbalanced data; proposed actionable insights to improve customer retention.
- Tools used: Python, Pandas, Scikit-learn, Matplotlib, Seaborn.

Student Mentor – Machine Learning Scientist Microso Student Club, Kafr El-Sheikh
Feb 2025 – May 2025

- Guided undergraduate students in mastering machine learning fundamentals through structured lessons and hands-on projects.
- Led workshops and coding sessions on supervised learning, deep learning, and model evaluation.
- Mentored teams on real-world projects including classification, NLP, and computer vision using Python, scikit-learn, and PyTorch.
- Reviewed code and model designs to encourage best practices in reproducibility, version control, and explainability.

LANGUAGES

Arabic (Native), English (C1), Italian (A2)