Sanket Sonu

Email: sanketsonu3@gmail.com | **Mob:** +91-8553823657
Availability: Immediate | **Current Location:** India

Portfolio
LinkedIn: Sanket Sonu | LinkedIn
GitHub: Sanket Sonu | GitHub
Kaggle: Sanket Sonu | Kaggle

Summary

- 4 years 10 months of Industry Experience with 3 years 8 months of experience with **AI/ML Algorithms** and open to relocate.
- Working Experience in **NLP** & Extensive knowledge of **Python** with libraries such as **Sklearn**, **TensorFlow**, **Keras**, **PyTorch**, **NLTK**, **SpaCy**, **OpenCV**, NumPy, Pandas, Matplotlib, & Seaborn. Worked on tools like - PyCharm, Jupyter Notebook, VS Code, **Docker, AWS SageMaker** & **GCP Vertex AI**.

Education

National College of Ireland, Dublin **Jan 2021 – Jan 2022**
MSc in Data Analytics | Grade: First Class Honours (1.1)

Sapthagiri College of Engineering, Bangalore, India **Jul 2012 – Jul 2016**
Bachelor of Engineering in Information Science & Technology

Skills

- Python • Statistics • Machine Learning • Deep Learning • TensorFlow • Keras • PyTorch • REST API
- Natural Language Processing (NLP) • Social Media Analytics • Data Visualisation • Agile • CUDA
- Google Cloud Platform (GCP) – AutoML • GCP Vertex AI • Tableau • Power BI • Scikit-learn (Sklearn)
- SQL • MongoDB • Relation Extraction • NLTK • SpaCy • NLU • OpenCV • IBM SPSS • Docker
- AWS SageMaker • AWS S3 & RDS • NumPy • Pandas • Business Intelligence (BI) • A/B Testing • Jira

Publication – Research Thesis Project

LREC International Conference Jun 2022 - Marseille, France
Identifying Emotion for Code Mixed Hindi-English Tweets | Paper | GitHub
Book-title: Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference
Publisher: European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

- Pulled real-world data using Twitter's official REST API - 'Tweepy' which is of 9,165 manually annotated bilingual Code-Mixed Hindi-English tweets.
- Dataset contains 7 classes of emotions: **Happy, Sad, Angry, Fear, Disgust, Surprise,** or **No emotions**. Extracted features using a few Vectorizers and Word Embeddings.
- Introduced **SVC, Multinomial Naïve-Bayes, Logistic Regression, Random Forest, LSTM,** and **BERT** models with different hyper-tuning parameters and achieved a maximum of 74% accuracy.
- Invented 7 setups, which include different feature selection processes, and trained all the models for each setup, to find the best one.

Work Experience

Data Scientist (NLP/NLU) - Intern, Orcawise, Dublin, Ireland — Jul 2022 – Mar 2023

- Optimising & data processing before building a custom classification model for data quality.
- Used **Data Engineering** concepts for processing unstructured **RAW** text data & performed **Annotation**.
- Designed & improved complex custom models on top of pre-trained models using **BERT** & **LSTM** and performed **A/B testing.** Data Mining, problem-solving & attention to detail.
- Development of **NER (Named Entity Recognition), Coreference Resolution**, & **Relation Extraction** for articles using **SpaCy** & **BERT** models on real world data like tourism, articles, & user's feedback.
- Delivering **data-driven** actionable insights using **Knowledge Graph** & research on **NLP** techniques.
- Design, deliver, document, and presentation of user-friendly dashboards and reports.

Senior Game Data Test Engineer (AI/ML), Pole to Win International, Hyderabad, India
Jan 2021 – Apr 2021

- Performed **Machine Learning** methodologies to utilise game data for player ranking system (matchmaking of online players using rank and in-game behaviour), functionality, and rewards system.
- Performed **EDA** and Segmented data by creating Clusters using the **K-Means** algorithm.
- Classified player's feedback using **Machine** and **Deep Learning Classification** models like – **Random Forest, XGBoost, Naïve Bayes, LSTM, CNN,** & **Transformers** to improve the gaming experience.
- Verbal communication with business stakeholders & clearly and concisely explained advanced & complex analytical findings to non-analytical peers and business leaders.

Game Data Tester (AI/ML), Ubisoft Entertainment India Pvt. Ltd, Pune, India
Jun 2018 - Dec 2020

- Development of advanced **Machine Learning** techniques to utilise game data for player ranking system (matchmaking of online players using rank and in-game behaviour).
- Classified player's feedback and chats using **Machine** and **Deep Learning Classification** models & Applied NLP & worked closely with the Software Engineering team to create **Chat Filters**.
- Processing the raw data using **Data Engineering** Concepts. Performed **Annotation**, **Model Selection**, and **A/B testing**. Applied analytical skills & knowledge transfer for data management.
- Continually research new methods and technologies in the insights and analytics space, including **AI** and **Machine Learning** tools and techniques.
- Diagnosed and restructured existing or new game designs by recommending unique, creative, and innovative ideas by collaborating as a **CO-DEV**.

QA Engineer, Sun Technology Integrators Pvt. Ltd, Bangalore, India, Mar 2017 - Jun 2018

- Performed automation testing to check the functionality of games using Python scripting on platforms like PlayStation and Xbox, following compliance to enhance performance and reported bugs in **Jira**.
- Re-designed many test cases to hit on the critical bugs and improved the end-user experience.

Technical Projects

Twitter Sentiment Analysis - Analytic Vidhya Hackathon – 70th Rank | GitHub

- The training data incorporate 31,962 tweets with labels as negative or positive. Obtained features using **TF-IDF** (1,2) unigrams and bigrams.
- Build a base model using **SVC, Multinomial NB, Logistic Regression,** and **Random Forest.**
- The model acquired an accuracy of 76.87% on the 17,197 unlabelled test-data and scored **70th** on the **Leaderboard** out of 1,280 users and 17,272 registered users in Hackathon – 2021.

Natural Language Processing with Disaster Tweets – 247th Rank | GitHub

- Build **SVC, Multinomial Naive Bayes, Logistic Regression, Random Forest,** and **BERT** models with different hyper-tuning parameters.
- Structured 7 setups, which include different feature selection processes, and trained all the models for each setup to achieve an accuracy of 80.29% and scored **247th** on the **Kaggle's Leaderboard**.

Urban Sound Classification using Neural Networks | GitHub

- The data sets consist of 8,732 labelled sound excerpts of urban sounds from 10 classes. Derived features using Librosa library.
- Build & Fine-Tunned **ANN, CNN 2D,** & **LSTM** models and visualized the comparison between 8 optimizers.

Database and Analytics Programming Project | GitHub

- Generated 3 Covid datasets (41,000 rows) using both **Kaggle** and **RapidAPI (REST API)**.
- Uploaded all datasets to **AWS S3** and operated 3 databases like **AWS RDS MySQL, MongoDB,** & **IBM Watson.**
- Pushed and pulled data through each database and analysed Covid impact using visualisation. Defined **dagster** open-source framework for implementation of **pipelines.**

Certificates

- Google Data Analytics Professional Certificate | Oct 2021 | **Certificate**
- The Data Science Course: Complete Data Science Bootcamp | Oct 2020 | **Certificate**