SIJIA HAN

Calgary, AB, T2K 5M7
Sijia.han@hotmail.com | 1(902)-220-5399 | LinkedIn | GitHub

SUMMARY

Passionate Machine Learning Developer/Data Scientist with 1+ years of professional experience and 5+ years in programming in software development, specializing in natural language processing. I drive innovation through data-driven insights, creating and implementing advanced models to unlock valuable information. I am committed to ongoing learning and staying at the forefront of my industry to assist organizations in adapting to the dynamic landscape.

SKILLS

Technical Skills:

- **Programming Languages:** Python, Java, R, C, SQL
- **Data Science**: Pandas, Numpy, MySQL, PostgreSQL, MongoDB, Matplotlib, Altair, Seaborn, Plotly, ggplot, Web Scraping, ETL, Data Analysis
- **Machine Learning/Deep Learning:** PyTorch, TensorFlow, SpaCy, Keras, Transformers, Scikit-learn, BERT, SVM, Logistic Regression, HuggingFace, Unsupervised Learning, Supervised Learning
- **Natural language Processing (NLP):** NLTK, Sentiment Analysis, NLU, Language Detection, Information Retrieval, Intent Classification, Word Embedding (BERT, BoW, Tf-Idf), NER, GPT, ASR, Topic Modeling, Dialogue System (Chatbot), ElasticSearch, LangChain, Prompt Engineering
- **Tools:** Git, Linux, REST API, Docker, Jupyter Notebook, FastAPI, MTurk, Large Language Models(LLMs), Azure Cloud, RStudio, Streamlit, Azure Machine Learning Studio**,** Azure From Recognizer, Azure Cognitive Search, Jira, Excel, PowerPoint

Soft Skills:

- Adaptability: Learned the new technology stack the team required for the upcoming project in a limited time.
- Collaborative: Collaborated with technical and business team members to successfully finish the project.
- Communication: Demonstrated through the ability to convey complex technical information to a non-technical audience.

Languages:

- Mandarin (Native), English (Intermediate)

WORK EXPERIENCE

AltaML Calgary, AB

Associate Machine Learning Developer (Full Time Contract) May 2023 - Sep 2023

- Developed an end-to-end project integrating **Azure Form Recognizer**, **Azure Cognitive Search** and **Large Language Model (LLM)** and deployed by applying **Streamlit App**.
- Developed and designed a modular **Named Entity Recognition (NER)** data extraction pipeline using **Python**, which integrated **Azure Machine Learning Studio** and LLM to process large volumes of data efficiently.
- Developed **data processing** steps for pre-processing to clean the input datasets and post-processing to refine and enrich the NER results of the LLM responses.
- Applied LLM based on the **LangChain** framework to perform NER tasks on complex datasets with more than 5,000 rows of different domain data, which reduced the time to perform NER tasks by 70% and improved NER accuracy by 50% compared with the old method.
- Utilized **prompt engineering** skills to generate effective LLM prompts based on different tasks, improving the efficiency and accuracy of LLM.
- Designed a hallucination logic to overcome the hallucination of LLMs and increase the accuracy of LLMs.
- Conducted experiments and analysis on various data and methods and reported the results of data analysis to the team leader.

Tools/Technologies Used: NER, Python, LLM, Azure Machine Learning Studio**,** Azure From Recognizer, Azure Cognitive Search, Streamlit App, LangChain, Prompt Engineering,

Heyday by Hootsuite Vancouver, BC

NLP Data Scientist (Full Time Contract) May 2022 - July 2022

- Developed a Negation-Aware **Named Entity Recognition** (NER) system for chatbots in the E-Commerce domain and deployed with a user-friendly interface using **FastAPI**.
- Implemented a **zero-shot learning** system to identify products and attributes with implied negations in the utterances of user requests in informal English.
- Collected and constructed large-scale NER and negation scope datasets/knowledge bases from the target product domains extracted from Google Taxonomy and open data resources, such as SOCC.
- Built and optimized the **semi-supervised CRF NER** and **Rule-based** models to achieve 88.62% overall accuracy and 68.49% f1-score on the Heyday-provided validation datasets.
- Delivered weekly reports, presentations, and the final project report to the supervisors, technical and non-technical.

Tools/Technologies Used: Python, NER, FastAPI, Semi-supervised Learning, Rule-based Learning

EDUCATION

University of British Columbia Vancouver, BC
Master of Data Science in Computational Linguistics Sep 2021 - Nov 2022
Dalhousie University Halifax, NS
Bachelor of Computer Science with a Minor in Mathematics Jan 2017 - Seq 2021

- Sexton Scholar (2018), Dean's List (2019), Dalhousie in-Course Scholarship (2019)

PROJECT EXPERIENCES

AltaML 2023 Sustainability Hackathon Aug 2023

- Built an end-to-end Waste Detection and Disposal Recommendation project integrated Image Segmentation, Image-to-text, Large Language Model and Streamlit App.
- Applied Beautiful Soup for web scraping to extract text from over 450 government websites to generate a knowledge base.
- Utilized the FastSAM CNN model to generate the crop images, then generated the text descriptions for images using the BLIP model.
- Utilized OpenAI's Language Model to create detailed and specific instructions on how to recycle each identified object based on the object descriptions and our comprehensive knowledge database.
- The project won first place in the AltaML 2023 Sustainability Hackathon.

Tools/Technologies Used: Computer Vision, Web Scraping, Image Segmentation, Large Language Model, Streamlit, Image-to-text Generation

Spotify Popularity Prediction Aug 2022

- Built an ensemble model of Ridge, Random Forest, XGBoost, LightGBM, and Catboost with the Spotify dataset to determine the popularity of songs and studied feature importance using SHAP.
- Visualized the data features using Matplotlib and Seaborn.

Tools/Technologies Used: Numpy, Pandas, Scikit-learn, Seaborn, Matplotlib, Feature Generation, Feature Importance, SHAP

COVID Sentiment Analysis in Tweets Mar 2022 - Apr 2022

- Extracted topic-relevant tweet data through the Twitter API and processed, cleaned, annotated and analyzed the raw data collected.
- Achieved an accuracy score of 0.72 with the best model among the models of CNN+BiLSTM, BERTweet and Fine-tuned BERTweet and Interpreted each model with visualizations.

Tools/Technologies Used: PyTorch, TensorFlow, Matplotlib, Scikit-learn, Sentiment Analysis, Pandas, Transformers, Text Annotation, CUDA, NLP

Linear Regression Models on Facebook Dataset using R Dec 2021

- Built single and multiple (Additive and Interaction) Linear Regression Models on the Facebook dataset, deployed with GGplot and Plotly.
- Used hypothesis testing to derive inferences.

Tools/Technologies Used: R, ggplot, Plotly, Tidyverse, Hypothesis Testing, Statistics, Quantitative Analysis, Bootstrapping