Yi Han

Want to save some time? Personal website: https://jadehanyi.com/
(781) 922-6520 | han.yi1@northeastern.edu
https://www.linkedin.com/in/yi-han-2900a1137/ | https://github.com/hanyidaxia?tab=repositories

EDUCATION

Northeastern University Boston, MA
Ph.D. in Industrial Engineering(NLP directed) Dec. 2023

Relevant courses: Deep Learning, Natural Language Processing, Machine Learning, Algorithms, Advanced Prob & Stats

Northeastern University Boston, MA
M.S. in Data Analytics Engineering Dec. 2018

Relevant courses: Data Mining, Database Design, Data Visualization, Probabilistic Operation Research

SKILLS & AWARDS

Programming:	Python, R, C++, JAVA, Tableau, JavaScript, HTML, React, CSS
LLM-related:	Language Model design, Fine-tuning, RAG, DPO, PPO, KTO, RLHF, Agent design, COT, LORA, LongChain, Hugging face
Modeling:	Pytorch, Sklearn, Tensorflow
Database & Cloud:	AWS, Azure, SQL, SAS, SPSS
Graphic Modeling:	Gephi, Anychart, AUTO-CAD, POWEBI, FLUENT, GAMIT
Certificate:	Mooc certificates of ML (2020); DL on Coursera (2020)
Coding Awards:	IDETC-CIE 2022 Hackathon, Problem 3, 3rd place
Patent:	Multi-Modal Data-Driven Design Concept Evaluator

WORK EXPERIENCE

Simplisafe **Boston, MA**
Machine Learning Scientist Feb.2025-Now

CV-related work

Flagship Pioneering FL100 **Cambridge, MA**
Machine Learning Scientist Jan. 2024-Jan. 2025

- Machine Learning part Leading issue: Alleviating the LLM hallucination through social media **RAG**
  - Engineered a sophisticated **multi-agent system**
    - Designed and implemented a distributed ML pipeline integrating multiple foundation models (OpenAI GPT, Llama, DALL-E 2) with async processing
    - Architected a robust type-safe system using Pydantic for structured data validation and Instructor for reliable model outputs
    - Achieved high-performance parallel processing with 5-second latency for complex multi-modal generation tasks (text + images)
    - Implemented genetic algorithm
    - Built a flexible architecture supporting both single-agent and multi-agent graph configurations for different use cases
  - Architected and implemented a **scalable Vector Search RAG** system on AWS, leveraging:
    - Engineered an efficient ETL pipeline processing millions of Amazon product reviews and metadata
    - Designed and optimized vector embeddings for 90M+ products using advanced ML models
    - Achieved sub-3 second query latency for complex vector similarity searches with dynamic filtering
    - Implemented cost-effective infrastructure automation, maintaining monthly AWS costs under $1.2K
    - Demonstrated expertise in large-scale data processing and ML system optimization while keeping total implementation costs under $2K"
  - Engineered a high-performance zero-shot hierarchical classification system:
    - Developed a sophisticated multi-level text classification model handling 1,700 nested product categories with 4-8 hierarchy levels
    - Achieved 83% F1 score on complex hierarchical classification without traditional training data,

demonstrating advanced prompt engineering
        - Implemented hierarchical classification logic to handle deep category trees while maintaining high accuracy
    - Architected advanced social media analytics pipelines for trend prediction and market research:
        - Engineered secure data pipelines for TikTok and Reddit with OAuth compliance and robust data governance
        - Developed ML models for social trend prediction using multimodal data (video transcripts, user engagement, reviews) from TikTok
        - Implemented NLP-powered consumer insight generation system similar to IdeaApe
        - Built scalable ETL workflows handling real-time social media data streams
        - Designed automated insight extraction systems for consumer painpoint identification and trend analysis
- Software developing part
    - Architected and implemented RESTful microservices using Python, Flask-RestX, and FastAPI for high-performance API development
    - Engineered scalable backend services supporting real-time social media data processing
    - Collaborated with frontend team to design and optimize API contracts and data flow patterns
- Database part
    - Architected and implemented enterprise-scale ETL pipeline for Amazon product data:
        - Designed and built high-performance ETL workflows processing massive Amazon product and review datasets
        - Engineered scalable data architecture using AWS DocumentDB (MongoDB-compatible) for efficient NoSQL operations
        - Developed robust data validation and transformation pipelines ensuring data quality and consistency
        - Optimized database performance through proper indexing and query optimization strategies
- Consultant part
    - Consult for two CPG product companies and one health product company with the agentic system
        - Competitor analysis with the Amazon data pipeline
        - Price analysis with the AI agentic workflow
- Some one day project
    - Finetuned a BERT on AWS sagemaker using some GPT generated synthetic dataset for a classification task
    - Build a web crawler based on AWS Lambda and AWS proxy services

Lira: https://liraglobal.com/ **NC**
AI Tech Lead Dec. 2022 - Jun. 2023
Task Introduction in short words: **Reading lip moves and translate to text from silent video.**

- Data Science Part Lead the lip reading project
    - Designed several deep learning models for the lip-reading task, including Vision-Transformer-backbone-based, ResNet-backbone-based, and several pure-video based
    - Build the model, in the latest work, and subtraction based transformer-backbone model achieved the state of the art performance in the lip reading task
    - Deploy the model to the app platform (in process)
- Data Engineering Part Lead the database building and data annotation:
    - Co-Design the web for collecting lip reading dataset
    - Co-Build and manage the **Azure** database for the collected data

Merck Cambridge, MA
Data Scientist Intern Jan. 2022 - Jul. 2022
Task Introduction in short words: **Identifying protein related findings in documents to specific disease.**

- Built target liability assessment text analyzing model (includes a **search algorithm** and a **text classifier**) based on the paper of target (compound) searching results from PubMed
    - Capabilities of the model:
        - Extracting all the sentences related to the compound and customizable disease or symptoms
- Built front-to-end prototype of the **deep-learning** based compound analyzing model (**DCM**) for Merck historical compound pdf-format-based reports
    - Capabilities of the model:
        - Parse all the pdf and word files, extracting different sections from those files including the abstract, conclusion, and result
        - Perform the **NER** (name entity recognition) task from the extracted sections

BaiRong Financial Information Service Company Beijing, China
Machine Learning Intern Jun. 2018 - Sep. 2018
Task Introduction in short words: **Regression model with XGBoost and LightGBM**

- Created model for risk control through **Logistic regression** and **Stepwise regression** via R and Python; "bad customer" ratio decreased by 5.67%, and payment received ratio increased by 20.3%
- Stacked multiple **XGboost** into single model, compared with **LightGBM**; increased AUC from 0.68 to 0.76
- Feature engineering with real consumer data, integrated into XGBoost

RESEARCH

Ph.D. Research

Main Research: NLP methods for Consumer Needs Elicitation

Objective: Utilizing NLP method to Extract latent customer needs from customer reviews

Aspect Category Sentiment Opinion Extraction for Latent Needs Elicitation

- Created a **unified deep-learning sequence-to-sequence model** with the new position encoding and Loss which **BEAT** pure **T5** on our dataset
- Built a new position encoding algorithm for full size transformer, a new loss function incorporates cross entropy and KL divergence for multi-label, weight-wised task

May. 2022 – Feb. 2023

- Built an annotated dataset, including labels A (aspect), C (category), O (opinion), S (sentiment) and I (opinion implicit indicator)
- Created a **unified deep-learning sequence-to-sequence model** to extract all labels parallelly based on **T5**
- Conducted clustering analysis for opinion and aspect, identified most contradictory opinion with same aspect

Aspect-Sentiment Guided Opinion Summarization for Latent Needs Elicitation

- Built a sequence-to-sequence MAS-T5 model for the aspect and sentiment-oriented summarization of reviews with Pytorch
- Designed a hierarchical **max-pooling model** MAS, which can predict the sentiment and aspect label in word, sentence, and review level
- Utilized the output from MAS to assemble a **synthetic** supervised summarization data, which can be used for abstract summarization task
- Fine-tuned the **T5** model with the synthetic data

Aspect, Opinion, Sentiment Extraction for Latent Needs Elicitation

- Built a **BERT-NER** model for elicitation of customer needs based on online reviews with Pytorch
- Designed a highly weighted loss function to resolve the extremely unbalanced dataset
- Labeled and assembled the output from the BERT-NER as user needs simulation
- Utilized **BLUE** score to evaluate the results of the needs
- Developed a web crawler to obtain source data
- Designed a double layer of **CNN** on top of BERT as a post-training parallel comparison

Sentiment and Opinion Extraction for Latent Needs Elicitation

- Built a **data crawler** to organize the original dataset
- Built a **product attribute lexicon** for further analysis (sneaker lexicon)
- Designed two types of algorithms for the **attribute level sentiment analysis**
- Conducted clustering analysis of the customer expression based on the sentiment analysis results

Other Research:

Algorithm Course Design in the College of Engineering for Northeastern University

- Designed course content based on two textbooks *Algorithm Design* and *Algorithms*
- Built the course example code and course quizzes
- Developed exercises based on textbooks
- Drew graph demonstration for classic algorithms like recursive

Unsupervised Attribute Clustering Analysis Based on Customer Reviews

- Filtered and clustered critical product attributes with product description via Pytorch
- Conducted clustering analysis based on filtered attributes instead of product

PUBLICATIONS

- **Journal Paper**
  - Han, Y. Moghaddam, M.(2024) Domain knowledge as attention fixer in Large Language Models *Journal of Engineering Design (*doi:10.1115/1.4067212*)*
  - Shi, J., & Yi, H.(2023). Aspect Guided Abstractive Summarization for Safety Concern Information Extration. *JCISE (Under review)*
  - Han, Y., Nanda, G., & Moghaddam, M. (2022). Attribute-Sentiment-Guided Summarization of User Opinions from Online Reviews. *J. Mech. Des.*, 1–41. doi: 10.1115/1.4055736
  - Han, Y., & Moghaddam, M. (2021). Eliciting Attribute-Level User Needs From Online Reviews With Deep Language Models and Information Extraction. *J. Mech. Des., 143*(6). doi: 10.1115/1.4048819
  - Han, Y., & Moghaddam, M. (2021). Analysis of sentiment expressions for user-centered design. *Expert Syst. Appl., 171*, 114604. doi: 10.1016/j.eswa.2021.114604
  - The feasibility study of fire emergency evacuation in the integrated transport system-Beijing south railway station, *China Chemical Trade* (ISSN:1674-5167)

- **Conference Paper**
  - Shoes-ACOSI: A Dataset for Aspect-Based Sentiment Analysis with Implicit Opinion Extraction **EMNLP** 2024
  - A Design Knowledge Guided Position Encoding Methodology for Implicit Need Identification From User Reviews, **IDETC/CIE,** 2023
  - A Priori: Design Knowledge in AI, **DesForm**, 2023
  - Extracting latent needs from online reviews through deep learning based language model, **ICED**, 2023
  - Aspect-Sentiment-Guided Opinion Summarization for User Need Elicitation From Online Reviews, **IDETC/CIE** 2022