ZHAOYI HOU (JOEY)

github.com/joeyhou joeyhou@seas.upenn.edu (858)729-8929

RESEARCH INTEREST

Natural Language Processing, Machine Learning, Cognitive Science
Commonsense Reasoning, Natural Language Inference, Narrative Understanding

EDUCATION

University of Pennsylvania *Aug 2021 - Present*
Data Science, Master of Science in Engineering

University of California, San Diego *Sep 2017 - Jun 2021*
Data Science, Bachelor of Science

PUBLICATION

- [ACL 2022] Xiaochen Kev Gao, **Zhaoyi Hou**, Yifei Ning, Jingbo Shang, Vish Krishnan. *Towards Comprehensive Patent Approval Predictions: Beyond Traditional Document Classification.*
- [Under Review] Caitlin A. Stamatis, Jonah Meyerhoff, Tingting Liu, **Zhaoyi Hou**, Garrick Sherman, Brenda L. Curtis, Lyle H. Ungar, David C. Mohr. *The association of language style matching in text messages with symptoms of affective psychopathologies.*

RESEARCH EXPERIENCE

Computer and Information Science (CIS) at Penn Engineering *May 2022 - Present*
NLP Researcher (advised by Prof. Chris Callison-Burch)
- Project 1: Building a **creative language generation pipeline** to automatically generate Dungeons&Dragons scripts and to track the state of entities within the game;
- Project 2: Building a **procedure understanding model** to generate alternative options in a given procedure script and to generate corresponding rationales based on commonsense.

Computer and Information Science (CIS) at Penn Engineering *Jan 2022 - Present*
NLP Researcher (advised by Prof. Lyle Ungar)
- Building a **conversation extraction module** extract lexical information from participants' text messages and analyze emotion mirroring in depressed participants.

Data Mining Lab at UCSD *Jun 2020 - Apr 2021*
NLP Researcher (advised by Prof. Jingbo Shang)
- Built a **text data ETL and classification pipeline** to handle **600,000** patent documents;
- Implemented a customized **BERT-based text classification model** and improved true negative rate (specificity) from **60% to 86%** (heavily unbalanced data with **84%** positive instance).

Salk Institute for Biological Studies *Jul 2019 - Jun 2021*
Data Science Researcher (advised by Prof. Satchidananda Panda)
- Built a **text processing pipeline** to extract food content from a health monitor app **(85% parsing accuracy)**;
- Built a **data ETL and analysis pipeline** for user behavior analysis (more than **500,000 records**).

PROJECT

Amazon Alexa TaskBot Competition *Nov 2021 - Apr 2022*
- Implemented the **document retrieval module** for the Alexa TaskBot competition;
- Improved the retrieval success rate by **25%** and advanced to the final.

Music Re-listen Prediction *Nov 2021 - Dec 2021*
- Built a **click-through-rate(CTR) prediction model** based on historical user behavior;
- Built a **feature engineering pipeline and deep factorization machine model** to achieve **67% AUC** in the un-seen test data.

Stack Overflow Question Quality Classification *May 2021 - Jul 2021*

- Built a **text data ETL and analysis pipeline** for **60,000** Stack Overflow question texts;
- Built an **XGBoost classification model** with AWS SageMaker and deployed it as an **AWS SageMaker Endpoint**;
- Achieved **87% accuracy** in "High-Quality Question" classification task.

TECHNICAL SKILL

Programming:	Python, Java, C, C++, Data Structure, Object Oriented Programming, SQL
Machine Learning:	LSTM, Transformer, CNN, Boosting, Logistic Regression, SVM, Decision Tree, K-Means
Software & Tools:	PyTorch, Hugging Face, AWS SageMaker, Pandas, Scikit-learn, MySQL
Data Analysis:	Bootstrap, ETL, Hypothesis Testing, Research Design