# Handan YU

handany@student.unimelb.edu.au | +61 0412716630 | HandanYU.github.io

## TECHNICAL SKILLS

- Proficient in processing and visualizing data using Python, specially Pandas, numpy and matplotlib packages
- Excels at MySQL to insert, delete, search, edit data from databases
- Familiar with R to do data analysis
- Extensive and in-depth knowledge of Machine Learning
- Know about feature selection algorithms in Machine Learning, like XGboost, Light GBM
- Be able to run on GPU under Linux and know about basic knowledge of Linux command-line
- Familiar with PyTorch and be able to use it to establish deep learning models
- Previous experience with git
- Familiar with mechanism of Hadoop, Spark; Skilled at transforming complex problems to mathematical modellings

## EDUCATION

**Master of Information Technology/Artificial Intelligence** **Jul 2021 - Jul 2023**
**The University of Melbourne**

- Got H1 in course Introduction of Machine Learning & NLP
- Achieve top 20% of competition in Kaggle named Build and evaluate sentiment classifier for tweets
- Achieve top 11% of competition in Kaggle named Rumour Detection

**Bachelor of Science/Information and computing science** **Sep 2017 - Jul 2021**
**Hangzhou Normal University**

- Graduated summa cum laude-4.28/5.0 GPA
- Major courses: Multivariate statistical regression, Data Mining, Big Data Analysis, Machine Learning, etc.
- Received First-class scholarship for outstanding students several times
- Won Honourable Mention Award in Mathematical Contest in Modelling
- Graduation project: research on sensitiveness of EM algorithm applying on clustering (https://github.com/HandanYU/EM_PROJET)

## RELEVANT EXPERIENCE

**Data Engineering Research Intern** **Jul 2022 - Present**
Walter and Eliza Hall Institute of Medical Research (WEHI), Melbourne, VIC, Australia

- Research data pre-processing methods to reduce the resources needed for cryo-EM

**Algorithm engineer Intern** **Feb 2022 - Jun 2022**
MindRank AI, Hangzhou, Zhejiang, China

- Via iteratively augmenting node features with gradient-based adversarial perturbations during training Graph neural network, the accuracy of molecular classification is improved around 2%.
- Improving traditional implementation of drop-out layer in Graph Neural Network by considering edges, nodes
- and even layer level helps the RMSE of Molecular property prediction decrease.
- Train and evaluate baseline models on benchmark.
- Reproduction latest paper related to Molecular Generation, Molecular property prediction.
- Utilize PyTorch framework to reproduce public projects implemented in Tensorflow

**Algorithm Engineer Intern** **May 2021 - Jun 2021**
GuanData Co., Ltd, Hangzhou, Zhejiang, China

- Assisted algorithm engineer to analyse daily output data using Pandas, write Error Analysis reports

- Collaborated with algorithm teams to cluster customers into small-customers or large-customers applying KMean algorithm. Designed and accomplished Kanban logics via MySQL
- Redesigned code architecture according to Pipeline to make it easy to review and debug
- Cooperated with algorithm teams to manage projects using git

**Database developer Intern** **Sep 2021 - Jan 2022**

Bosch Power Tools China Co., Ltd, Hangzhou, Zhejiang, China

- Collaborated with team for demand collection and architectural decisions
- Architected, designed, and developed management software's front-end and back-end utilizing pySimpleGUI tool to meet with demand of projects' capacity planning
- Reprocessed data using Pandas in Python to reduce redundancy of data
- Improved management efficiencies and implemented reasonable allocation of capacity

**Assistant of algorithm engineer** **Dec 2020 - Mar 2021**

HikRobot Technology Co., Ltd, Hangzhou, Zhejiang, China

- Created script using Halcon and pyzbar to tag QR code and barcode to reduce manual operation; increase efficiency of tagging by 30%
- Created and developed automation script for model training on VM software using pywinauto and pyautoGUI to remove manual execution; reduced total model training time by 20%
- Assisted deep learning team with collecting training datasets, training target detection algorithm models and performing data verification

## RELEVANT PROJECT WORK

**Kaggle Competition - Rumour Detection on tweets (Group)** **April 2022 - May 2022**

**GitHub: https://github.com/HandanYU/Rumour-detection**

***Construct Deep learning models based on context feature and statistics of tweets to predict the tweets are rumour or not and use data mining approaches and LDA models to analyze the characteristics between rumours and non-rumours.***

- Use Regular Expression and NLTK package to do pre-process (including removing stop words, lowercasing, stemming on tweets)
- Extract statistical features among raw tweet objects and develop some statics such as following rate
- Utilizing KNN interpenetration to fill missing data improves the F1 score on development dataset around 3%, compared with mean filling
- Implement deep learning models combined BERT with three layers full connected neural network.
- Finally, the model based on BERTweet pre-trained model with statistics selected by LightGBM achieved the best performance on test dataset (F1 score on rumours: 0.90797, top 20% in Kaggle competition)
- LDA model would be used to analyse the topics of rumours and non-rumours

**Kaggle Competition - Twitter Sensitive Analysis (Independent)** **Sept 2021 - Oct 2021**

**GitHub: https://github.com/HandanYU/Twitter-Sensitive-Analysis**

***Construct multi-classifiers based on machine learning models to predict the sensitive of the tweets according to TFIDF of the tweets' context and analyze the gender bias of the classifiers.***

- Implement data mining on the training dataset and put forwards baselines based on Zero-R and opinion lexicon
- Built three types of sentiment classifiers (namely Naive Bayes, Softmax Regression, and K Nearest Neighbor) to deal with a 3-way task of classifying tweet sentiment into positive, neutral, and negative
- Examine these classifiers separately on both traditional TFIDF and the improved TFIDF combined with information entropy feature selector. As for the result, improved TFIDF can improve the accuracy of all classifiers on this task
- Through constructing a special formula to discuss the impacts of gender bias on the performances of classifiers, it would be found that classifier based on NB was the most sensitive with the gender bias on training dataset

**The Fish Migration Problem in the Scotland Based on LSTM-RNN(Group)** **Feb 2020 - Feb 2020**

- Conducted research on migration of Scotland fish based on LSTM

- Built LSTM model to predict temperature of sea surface in coming 50 years. The MSE of final model is as low as 3.19 Fahrenheit degree.

### Taxi Assignment Problem in the Airport Based on Queue Theory(Group) Sep 2019 - Sep 2019

- Accomplished a research on taxi scheduling in airport based on queuing theory
- Pre-processed data and implemented feature selection
- Built reasonable taxi scheduling model according to real situation of Shanghai Pudong Airport to decrease waiting time of both customers and drivers