Haohua Chang

+86 181-8382-4122 • haohuachang@outlook.com

EDUCATION

Massey University Auckland, New Zealand
MSc Information Sciences 02/2023 - 11/2023
GPA：7.0/9.0 | Grade: Merit
Key Modules: Data Science - Making Sense of Data, User Interface Design and Evaluation, Location Systems Spatial Databases, Tools and Applications, Location Data Mapping, Analysis and Visualization.

Shanxi Agricultural University Shanxi, China
MSc Information Sciences 08/2023 - 12/2023
GPA：4.13/5.0 | Grade: Distinction
Key Modules: Natural Language Processing, Image Processing, Pattern Recognition, Advanced Databases.

Kunming University of Science and Technology Kunming, China
Bachelor of Engineering 09/2018 - 06/2022
GPA: 3.39/4.0
Key Modules: Data Structure, Operating System, Airborne Electronic Systems and Devices, Fundamentals of Control Engineering, Fundamentals of Electrical and Electronic Technology, Principles of Mechanics, Mechanical Design, and Fluid Mechanics and Aerodynamics.

RESEARCH EXPERIENCE

Project：Kaggle competition (LLM Prompt Recovery) China 2024
Research Background and Aims:

- NLP workflows increasingly involve rewriting text, and this competition helps to learn how to effectively cue LLMs
- The aim of the competition is to design algorithms to recover the Large Language Model (LLM) rewriting prompts used to rewrite a given text.
- The competition was tested on a dataset consisting of more than 1300 original texts and their rewritten versions in Gemma (Google's Large Language Model). The evaluation criteria are based on the embedding vectors generated by the sentence-t5-base model, and the score is calculated by Sharpened Cosine Similarity.

Content Abstract:

- Generate training data: Using Huggingface's Openwebtext to generate original text, using the gpt-3.5-turbo API to generate rewrite prompts, and then randomly pairing the original text with the rewrite prompts and inputting them into the LLM model gemma-7b-it to generate rewrite text.
- Data filtering and cleansing: removing text with a word length of more than 512 in the original text, and removing rewritten text that has no practical significance.
- Ensemble learning: Build three models (deberta-v3-large, Mistral-7B-instruction-V0.2 and SBT-Phi2) and concatenate their prediction results as the final prediction result.
- Train the Deberta model with the generated data, experimented with different hyperparameters of the Deberta model and added a linear layer during the fine-tuning process, experimented with the size of the linear layer. Use open-source model Mistral-7B-Instruct-v0.2, input examples (few-shot) for direct prediction and prune the prediction to remove redundant symbols or text. Employ an open-source Phi2 fine-tuning model for prediction, focusing on key text segments.

In conclusion, the deberta model performs best when the size of the linear layer added to the deberta model is 32768, and concatenating the predictions of the three models can effectively improve the performance.

Project：Disaster Impacts in Social Media: Classifier and Named Entity Recognition
MSc final research project (**individual** and part of the wider project QuakeText) Massey University
Supervisor: Kristin Stock 2023

Research Background and Aim:

- The large amount of real-time data from social media platforms at the time of a disaster can improve situational awareness of the disaster event. The goal of the Quake project is to use an API to obtain a large amount of real-time social media data, and then use a classifier to filter out disaster-related data. After that, a named entity recognition model is used to extract fine-grained disaster information with geographic locations, and finally display this information in a web map application.

- The aim of this project is to implement high-performance classifier and evaluate the quality of data from different social media platforms.

Content Abstract:

- This study compares the performance of classifiers based on keyword matching, machine learning (logistic regression, random forests, support vector machines, etc.), and deep learning (CNN, LSTM, transformer-based models, etc.) in this research area, and implements the best-performing classifier which is based on the Roberta model (with an F1 score of 0.871) to categorize the data into 11 categories and filtered the categories containing fine-grained disaster information based on their relevance to the disaster information.
- In this study, the performance of the classifier was validated using the human-labeled Kaikoura earthquake dataset, and the F1 score of the classifier on this dataset was 0.896.
- This study developed a data retrieval strategy and used Google APIs with web crawling techniques to obtain a large amount of data from five social media platforms including Youtube, Facebook, Instagram, Titok, and Reddit, and evaluated the availability, richness, and reliability of the data as well as the feasibility of obtaining the data.
- The study compares the proportion of relevant data and the proportion of relevant data containing location information (a classifier with NER model was applied to extract fine-grained information about disasters in the data) to evaluate the availability of the data, compares the total number of data, the number of relevant high-frequency topics and the number of irrelevant high-frequency topics (Thematic Analysis) to evaluate the richness of the data, and compares the proportion of error information to evaluate the reliability of the data. Finally this project compares the usability of APIs of different social media platforms.

In conclusion, this study implements a disaster data classifier that yield promising results. The data evaluation indicates that Facebook, Instagram, and YouTube have more available data, but Facebook is richer in data and better able to focus on specific disaster events. In addition, the API provided by Facebook and Instagram enables access to more real-time data in a short period of time. Therefore, Facebook is the best source of data for this project.

Project：Predicting the best player of a season in the NBA Massey University

Research Aim:

- The purpose of this project is to predict the Most Valuable Player (MVP) of the year for the NBA regular season by using machine learning algorithms

Content Abstract: 2023

- Data Acquisition: The game statistics of all NBA players in the past 43 years were acquired through web crawling technology and website API, including 34 technical data, i.e. 34 features, with a total of 16,605 samples.
- Feature engineering and algorithms: heat map was drawn to calculate variable correlation, feature importance was calculated using random forest algorithm, new features were created. Finally, 14 features were selected and predicted using knn algorithm with random forest algorithm.
- Evaluation and optimization: using random grid search (large scale) and grid search (small scale) to select optimal model parameters and adjust category weights to address data imbalances
- Front-end: This study designed a graphical user interface using Streamlit

In conclusion, this project has an AUC of 0.891 and predicts that the probabilities of the top five MVP candidates in 2023 are, in order, Giannis Antetokounmpo 0.28, Joel Embiid 0.23, Nikola Jokic 0.13, Jayson Tatum 0.11, and Domantas Sabonis 0.10. The MVP for 2023 has already been revealed as Joel -Embiid.

Project: Kaggle competition (IEEE-CIS Fraud Detection) Massey University 2023

Research Aim:

- The goal of the competition was to build models to predict whether a customer transaction was fraudulent or not and provided more than 735,000 samples and more than 300 features, including both categorical and numerical features.

Content Abstract:

- Data processing and analysis: filling in missing values, standardizing data, comparing positive and negative sample sizes, and analyzing the data distribution between the training and test sets
- Feature engineering: drawing heat maps (calculating variable correlations) and filtering out features with high correlation to the target variable and removing highly correlated features. Using the Random Forest algorithm to calculate feature importance.
- Experimental models: knn algorithm, plain Bayesian algorithm (mixed Gaussian and Gamma distributions), LightGBM classifier
- Evaluation and Optimization: this study uses cross-validation with multiple evaluation metrics (AUC, Accuracy, F1, Precision, Recall, Specificity, FPR) to validate model performance, uses grid search to select the optimal model parameters, and builds a small dataset (one-tenth of the size of the original dataset) in

order to implement more experiments in a short period of time. In addition, this study addresses the problem of positive and negative sample imbalance by setting a classification threshold.

In conclusion, this study focuses on the optimization of knn and plain Bayesian algorithms. Eight features are selected as inputs from hundreds of features given on the official Kaggle website. The optimized knn model achieves the best performance with an AUC of 0.853.

Project: Planning algorithm designs for colleges and universities China 2021-2022

Kunming University of Science and Technology

Bachelor final research project | **Score: 85.9/100**

Project Leader, Project Planning, and Implementation

Research Background and Aims:

- With the increasing richness and diversity of subject categories in universities and colleges, class scheduling problem, as a multi-constraint, multi-objective combinatorial optimisation problem, has been proved to be an NP-complete problem, and thus it is quite a complex and difficult task to achieve automated scheduling that is efficient, has reasonable scheduling schemes and can meet the special requirements of each school.
- The goal of this study is to design an efficient and reasonable scheduling algorithm with high adjustability to meet the scheduling needs of different universities.

Content Abstract:

- The project mainly uses the time matrix model to solve the time conflicts among classrooms, teachers and classes, and after solving the main time conflicts, it arranges the time and classrooms for the courses based on the idea of priority algorithm and greedy algorithm.
- The project classified all courses into four categories and designed algorithms for these four categories that were more appropriate for them based on a unified scheduling idea.This project sets teacher unavailability time prior to scheduling to meet the needs of teachers
- The optimal combination of time based on students' learning efficiency and teachers' preparation time improves the scheduling efficiency of courses with a high total number of hours and enables such courses to be organized in a more rational way.
- The project prioritizes available time slots by combining the learning efficiency of the slot with the total number of hours the teacher has to teach on that day, so that lessons are evenly spread over the best of the available time slots and the teacher is not overloaded with lessons during the day.
- Developed automatic class scheduling software using C++ language. Input and output interfaces were designed to transfer data in the form of files. Finally developed GUI based on MFC

In conclusion, this project exemplifies the scheduling process for courses in a semester at the School of Civil Aviation and Aeronautics of Kunming University of Science and Technology. The scheduling test of the given 62 courses and 20 classes shows that the automatic scheduling software can fully realize the data import function, automatic scheduling function and scheduling result query function of the university scheduling system, and it also realizes the function of setting teachers' unavailability time specially designed in this project. The final teacher schedule, class schedule, classroom schedule results are clear and reasonable, and able to meet the general scheduling needs of the university.

PUBLICATIONS

Recent Progress in Development and Application of Proteins from Fish Viscera

EI & Chinese Science Citation Database (CSCD) | Funded by the National Natural Science Foundation of China & Natural Science Foundation of Guangdong Province

Author(s) :

Journal article :

DOI:

WORK EXPERIENCE

INTERNSHIP EXPERIENCE

AWARDS

Scholarships:

- Kaggle Competition Silver Medal (LLM Prompt Recovery) — China 2024
- Master's Scholarship Project (QuakeText) — New Zealand 2023
- Second Class Outstanding Student Scholarship at the University (first semester) — China 2021
- Second Class Outstanding Student Scholarship at the School of Civil Aviation and Aeronautics — China 2021
- Third Class Outstanding Student Scholarship at the University (second semester) — China 2020
- Third Class Outstanding Student Scholarship at the University (first semester) — China 2020
- Second Class Outstanding Student Scholarship at the University (first semester) — China 2019
- Third Class Outstanding Student Scholarship at the University (second semester) — China 2018

SKILLS AND INTERESTS

Languages:

- IELTS 7.0 (L:7.0, R:7.5, W:6.5, S:6.5)

Skills:

- Proficiency in data analysis using Python.
- Mastery of basic machine learning algorithms.
- Proficiency in programming with Python and C++ languages.
- Proficiency in using QGIS, PostGIS, Geoserver, OpenLayers.
- Proficiency in 3D mapping using Solid Works.

Interests:

- Basketball, badminton, table tennis, running