Yule Wang

yulemoon@gmail.com • +1 (604) 783-2901 • https://github.com/yulew • https://www.kaggle.com/moonswords

Education

- **Doctor of Philosophy, Physics,** 3.89/4.0 Sep. 2015 – July. 2021
  Simon Fraser University, Burnaby, BC, Canada
- **Master of Science, Physics,** 3.67/4.0 Jan. 2013 – Aug. 2015
  Simon Fraser University, Burnaby, BC, Canada
- **Bachelor of Science, Applied Physics,** 86/100 Sep. 2008 – Aug. 2012
  Harbin University of Science and Technology, Harbin, China

Work Experience

- **Applied Quantitative Methods (AQM)** – Best Buy Twitter Data Jan. 2017 - Nov. 2017
  Data Analyst, intern *Vancouver, Canada*

· Led the cross team NLP project of Best Buy Twitter data between AQM and Best Buy.

· Automated a pipeline that transformed the data crawled from Twitter API into JSON format structural data. Cleaned, parsed and segmented the Twitter texts (NLTK, word2vec and TfidfVectorizer).

· Built a content-based spam tweets filtering system, reaching 90% accuracy, by detecting duplicate tweets from different accounts and applying key-words *Naive Bayesian classifier*. It corrected overall positive sentiment score of the original dataset by -11%.

· Established the sentiment analysis to evaluate the satisfaction improvement of customers after being responded to by Best Buy customer service on Twitter using the *support vector machines* (SVM) model;

· Successfully classified topics of Best Buy tweets using the *Latent Dirichlet Allocation* model;

· Performed gender classifications of Twitter usernames, and achieve beyond 90% accuracy, by implementing the unsupervised character n-grams algorithm. Found different sentiment improvements for males and females.

Kaggle Competitions 2018 - 2020

- **TalkingData AdTracking Fraud Detection Challenge** *bronze medal, top* 8% - leader Mar. 2018 - May. 2018

· Led a team of three and won a **bronze metal** in the competition of detecting fraudulent clicks on mobile advertisements.

· Improved 0.8% of AUC-ROC score by oversampling/downsampling the highly imbalanced dataset (SMOTE, negative-sampling).

· Implemented the baseline LightGBM model and improved 0.6% of AUC-ROC score by ensembling a fully-connected neural network.

· Engineered 50 time-series group-by features (combinations of IP, OS, APP, channel, day, hour, next click interval, etc.). Improved 2% of AUC-ROC score by selected top 30 important features (ranked by LightGBM).

- **Deepfake Detection Challenge** Mar. 2020 - Apr. 2020

· Built a pipeline to detect deepfaked videos which successfully handled large volume of videos (500 GB training data).

· This pipeline automated the process of extracting face embeddings from frames in videos using facenet-pytorch and feeding the face features into sequential neural network (convolutional LSTM) within the limited RAM and SSD resources.

- **Toxic Comment Classification Challenge** Jan. 2018 - Mar. 2018

· Led the team in the competition of classification of different types of toxic sentiments of comments in the Wikipedia's talk page.

· Improved 0.8 % of AUC-ROC score by carefully preprocessing texts including words corrections and foreign language translations.

· Vectorized words using pretrained GloVe embeddings. Fed the word-embeddings into GRU/LSTM models and improved 0.3 % of AUC-ROC score by ensembling these models.

Research Projects

- **PhD: Statistical Modelling and Simulations of Failure Dynamics in Random Networks** Sep. 2015 – Dec. 2020

· Built a statistical prototype model on a random graph that quantitatively described the random failure *dynamics* in polymer *networks*. Successfully forecast polymer networks' failure times which matched the real world polymers' lifetimes.

· Established a kinetic *Monte Carlo* program from scratch in *Python* that successfully simulated the *Markov* fracture processes.

· Optimized and implemented an efficient algorithm of locating the connected clusters on the graph (percolation theory) that improved the time complexity from $O(N^2)$ to $O(N)$.

· Successfully conducted the large-scale simulation on cloud and improved the RAM efficiency by reducing the redundant usage that occurred during the huge matrices computation.

Programming/Data Skills

- **Languages**: **Python** (6 years), SQL, Bash, Git, R. **Libraries**: Keras, Scikit-learn, Numpy, Pandas, Matplotlib, NLTK.
  Skills: statistical modeling, Random Forest, Boosting (XGBoost/LightGBM), NLP: word embedding (word2vec), LDA, LSTM.

Selected Publication

- Wang, Y. and Eikerling, M. Fracture dynamics of correlated percolation on ionomer networks. Physical Review E **101**, 042603 (2020). (https://journals.aps.org/pre/abstract/10.1103/PhysRevE.101.042603)