Tianyi Sun

Tel: 612-309-0898 | E-mail: sun00234@umn.edu | Web: https://tianyisun00234.github.io/

EDUCATION

University of Minnesota Twin Cities   **Aug. 2018 – May 2021 (Expected)**

Bachelor of Mathematics (Computer Specialization), Minor in Statistics and Computer Science   GPA: 3.67/4.0

Relevant coursework **(upper-division and graduate courses)**: Artificial Intelligence, Machine Learning, Algorithm and Data Structure, Formal Language and Automata, Probability and Statistics, Applied Linear Algebra, Numerical Method, Cryptology and Number Theory, Mathematical Logic, Theory of Statistics, and Regression and Correlated Data.

Central University of Finance and Economics, Beijing   **Sep. 2016 – Jun. 2018**

Mathematics   GPA: 3.84/4.0 (top5%)

Relevant coursework: Linear algebra, Calculus, Macroeconomics, Microeconomics, Fundamentals of Accounting, Public Finance, Fundamentals of Statistics, Business Statistics, Marketing Management, Psychology, and Sociology.

HONORS & SCHOLARSHIP

- University of Minnesota's Undergraduate Research Opportunities Program (USD $1,500)   Spring 2021
- Maroon Global Excellence Scholarship (USD $15,000)   Fall 2018 – Fall 2021
- Vice president candidate of Tau Sigma National Honor Society Uni. of Minnesota Twin Cities Chapter   Fall 2020
- Membership of Tau Sigma National Honor Society Uni. of Minnesota Twin Cities Chapter   Spring 2019 – Present
- Dean's List of College of Library Arts at the University of Minnesota   Spring 2019 – Present

RESEARCH INTEREST

My research interests lie in the general area of machine learning, specifically in deep learning, unsupervised learning, few-shot learning and their applications in natural language processing and sequential decision making, including GPT-3, Natural Language Generation and Natural Language Understanding. I am also interested in applying AI techniques to address societal challenges such as combating the COVID-19 pandemic.

RESEARCH EXPERIENCE

Improve Natural Language Understanding   **Oct. 2020 – Present**

Independent Researcher, Advisor: Prof. Maria Gini.

- Optimize language pre-training models e.g. BERT and RoBERTa.
- Design a model combining logical reasoning, Bayesian Deep Learning and language pre-training model.
- Implement Few-Shot Learning to generate texts on the dataset of emotion responds to COVID-19.
- **Grants supported by the University of Minnesota's Office of Undergraduate Research.**

How personal perceptions of COVID-19 have changed over time   **Jul. 2020 – Sep. 2020**

Independent Researcher, Advisor: Prof. Maria Gini.

- Aimed to analyze personal perceptions towards the COVID-19 pandemic with the main challenge emanating from the limited amount of data and paucity of previous works.
- Proposed a perception analysis method combining sentiment analysis with topic extraction and sequential prediction, discovering the first ground truth COVID-19 emotion responds dataset at ACL-2020.
- Designed a model evaluation scheme to select the most optimal one for sentiment analysis using Naïve Bayes, Random Forests, Linear SVM, Logistic Regression, LSTM, BERT, RoBERTa, and DistilBERT.
- Extracted five topics from the first ground truth dataset using LDA.
- Made sequential prediction of trends in five topics and thirteen sentiments using ARIMA and Encoder-Decoder LSTMs.
- Estimated the health status of users in Reddit and discovered users' consistent nervousness about COVID-19.
- **Paper submitted at AAAI2021. Presented in the Chatbot Group at UMNTC.**

Clustering U.S. counties to find patterns for COVID-19 pandemic   **Jul. 2020 – Sep. 2020**

Group member of Ecolab-UMN Collaboration, Leader: Sarah Milstein.

- Discovered patterns relating to the COVID-19 pandemic for each U.S county and found the core transition factors.
- Constructed a dataset of data relevant to the spread of COVID-19 from WHO and Johns Hopkins University.
- Implemented and evaluated K-Means, Fuzzy c-Means, Mini Batch K-Means, Gaussian Mixture Models, and tuned their hyperparameters using appropriate methods choosing from Silhouette Metrics, Calinski- Harabasz Index, Davies-Bouldin Index, Elbow Score, AIC, and BIC.
- Optimized clustering interpretation method using Jenks Natural Breaks Optimization and Decision Tree.

- **Paper submitted at SIAM.**

Forecasting daily COVID-19 spread in regions around the world   **Mar. 2020 – Jul. 2020**
Group member of Ecolab-UMN Collaboration, Leader: Jimmy Broomfield.
- Predicted the COVID-19 confirmed cases and fatalities for each region and country around the world.
- Implemented Encoder-Decoder Bidirectional LSTM, ARIMA with Square Root Transform, ARIMA with Log Transform, Multiphase Logistic Model, and Fill Forward Model to select the optimal one for prediction of each region.
- Optimized epidemiological SIR model into SEEAIRD model using incubation, infections not yet identified, asymptomatic carriers, and death compartments, which improved our rank by 31 places, to 35th out of 250+ in Kaggle.
- Derived ordinary differential equation, transition probabilities, system of stochastic differential equations, numerical simulation and parameters estimation of SEEAIRD model.
- **Paper is currently in the process of preparation.**

FDA COVID-19 Risk Factor Modeling Challenge   **Jun. 2020 – Jul. 2020**
Group member of Ecolab-UMN Collaboration, Leader: Jimmy Broomfield.
- Investigated how race, ethnicity, age, and history of comorbidities affect the progress of COVID-19 infected veterans.
- Found risk factor and protective factor: a number of COVID-19 infections died due to the history of chronic comorbidities instead of COVID-19; infections who got PCV vaccines are less likely to die from COVID-19.
- Made the final prediction of Alive or Deceased Status by integrating veterans' comorbidities records and COVID-19 Status, Days in ICU, and Controlled Ventilation Status.
- Discovered the inconsistency of categorical values between training and test sets and proposed a strategy of transforming the values in training set to match the ones in test set, which significantly improved the modeling accuracy.

MUDAC 2020: Investigating Disparities in Outcomes across Venues   **Mar. 28th – 29th, 2020**
Data Science Challenge participant, Advisor: Prof. Gilad Lerman.
- Discovered count venues' tendency to favor the plaintiffs or defendants and predicted the probability that a case will be closed by a summary judgment.
- Evaluated the performance of Logistic Regression, Support Vector Machine, Decision Tree and Random Forests for prediction, using Random Forest Feature Selection to improve modeling accuracy by more than 30 percent on average.

INTERNSHIP EXPERIENCE

CenterPoint Energy   **Apr. 2019 – Mar. 2020**
Data Analyst, *Joblogic-X Corporation,* Supervisor: Tengran Liu.
- Optimized the customer entry methods by designing a model to automatically duplicate the entry context into another cell, largely saving customer's entry time and being used in other projects.
- Developed SSIS (SQL Server Integration Service) data flow to ingest data from various sources and leveraged the SSIS source reader to process flat files, XML documents, and other related sources.
- Designed standard data quality routine to clean the source data and keep track of data quality matrixes.
- Time series predicted the products' weekly inventory from suppliers and created reports through SAP Business Objects.
- Reached out to suppliers if the information was unclear and sought opportunities to develop long-term cooperation.
- Analyzed prices, promotions, distances, delivery time, and qualities of suppliers to design optimized purchasing solutions for customers based on different types of products and customer requirements.
- Received return offer.

AWARDS & LEADERSHIP EXPERIENCES & VOLUNTEER EXPERIENCES

- 2nd Place in National Collegiate DanceSport Championships, Amateur Championship Latin Dance J/C/R   Spring 2019
- 2nd Place in Dance Fest, Cody Arndston Amateur Silver International Latin Dance S/C/R   Spring 2019
- Vice President of the Central Uni. of Finance and Econ. Students Union   Spring 2016 – Spring 2018
- Communication Coordinator of Beijing Daxing district No.1 middle school   Fall 2014 – Spring 2016
- Volunteer English teacher at Beijing No.2 primary school and in Galle, Sri Lanka   Winter 2018 – Fall 2015
- Passed the Gold Level Latin Dance examination (the highest) at Beijing Dance Association   Fall 2014

SKILLS

- Programming Language: Python, LaTeX, R-studio, Matlab, and SQL.
- Tools: SciPy, Scikit-learn, PyTorch, Keras, fast.ai, TensorFlow, Transformers, and NLTK.
- Data analysis: dataset construction, text preprocessing, machine learning and deep learning modeling and tuning.
- BI Tools: SAP Business Objects, Tableau, and Power BI.