HAOMIN LIN

humaslin97@gatech.edu | (404)-433-0697 | https://www.linkedin.com/in/humaslin | Atlanta, GA

EDUCATION

Georgia Institute of Technology, Atlanta, GA — Expected May 2021
Master of Science in Computational Science and Engineering — GPA: 4.0
Tianjin University, Tianjin, China — Sep 2015 – July 2019
Bachelor of Engineering in Optoelectronic Information Science and Engineering (TJU & NKU Organized) — GPA: 3.6

TECHNICAL SKILLS

Areas of Expertise: Data Analysis, Data Visualization, Machine Learning, Natural Language Processing
Languages and Technology: Python, C, SQL, PHP, R, D3.js, Tableau, SciKit-Learn, PyTorch, Tensorflow, Spark, Hadoop

EXPERIENCE

Teaching Assistant, Georgia Institute of Technology — **Jan 2021 – Present**

- Instruct 20+ students out of 150 on data analysis techniques using **Spark** and **deep learning** frameworks such as PyTorch.
- Guide students through **NLP** course projects such as analyzing Twitter data with the BERT NLP model to classify news and personal stories and perform sentiment analysis. Address student questions / help debug programs.

Machine Learning Research Assistant, Georgia Institute of Technology, *Prof. Munmum De Choudhury* — **Sept 2020 – Present**

- Studied media portrayal of immigrants and the corresponding effects on society throughout history through data.
- Scraped 370,000+ newspaper articles from research database and 97,000+ online media websites via Webhose.io API in **Python**. Collected data of real-world events concerning immigrants from GDELT via **Google BigQuery**.
- Implemented **word2vec** model with NLP toolkits (**spaCy**) to extract key adjectives describing immigrants constructed by media.
- Ran **regression analysis** on transformations of immigrant sentiment and real-world events concerning immigrants. Through finer-grained analysis, study illustrated significant correlation during 1960s, 2000s of negative immigrant media portrayal.

Data Science Research Assistant, Northwestern University, *Prof. Ágnes Horvát* — **July 2020 – Sept 2020**

- Investigated how online media on COVID-19 differed across three platforms including news, blogs, and discussions. Summarized work in a paper to be published on *Journal of Quantitative Description: Digital Media*.
- Pre-processed over 3 million entries of data to reduce their features to a 120-day series with **Pandas** in Python.
- Independently identified method to extract info from platform behavior variation (posts/day, etc.) and computed correlation score variance based on time series by shifting daily news coverage forward by a few days (lag days).
- Applied windowed time lagged cross correlation analysis to extract further information from correlation score variance to present temporal variation in amount of COVID coverage/sentiment scores, indicating news as the leader in online discourse.
- Used heatmaps to illustrate growing positive sentiment towards COVID recovery, developed with **Tableau** and **Python**.
- Conducted cross-platform **ANOVA** and **post hoc Tukey** tests to validate linguistic disparities in different platforms.

Data Research Assistant, Indiana University Bloomington, *Prof. Xiaojing Liao* — **May 2020 – Aug 2020**

- Studied and developed knowledge graph to identify possible damages from vulnerabilities to distribute appropriate resources.
- Extracted information from 4,989 products, 2,943 vulnerabilities and additional information from structured/unstructured text data in 1,316 network security reports and transformed features of vulnerabilities into usable data for knowledge graph.
- Used **regular expression matching** and **POS tagging** to refine the results and then structured the data for storage.

Student Worker, Georgia Institute of Technology, Community & Digital Archives Project — **Jan. 2020 – May 2020**

- Developed a plugin in **PHP** with VADER sentiment analysis lexicon to flag negative comments in the database in the backend for administrators to review. Over 80% warnings in the test phase were deemed appropriately flagged.
- Trained a model based on **Transfer Learning** in **Tensorflow** to identify maps from images with accuracy of over 90%.

RELATED PROJECTS

Automatic Options Trading Generation Via Distributed Deep Reinforcement Learning

- Designed application to advise on options trading actions based on **Q-learning** in reinforcement learning.
- Developed the model with **PyTorch**, trained by historical option prices, achieving up to 10% revenue per month.

Revealing Gendered Language in Job Descriptions

- Scraped over 300,000 text data from Indeed.com to investigate usage of gender languages in online job postings.
- Trained classifiers with **Ensemble Learning** in **Scikit-Learn** to predict salary levels of jobs with 87% accuracy.
- Categorized job descriptions into 16 industries by counting the frequencies of keywords from each industry.
- Visualized the usage of gender language in **Tableau** to present the difference between industries and salary levels, identifying clear gender biases across industries, specifically jobs in IT/finance or high-paying jobs.

PUBLICATIONS

- Dambanemuya, H.K., **Lin, H.,** and Horvát, E-Á. Characterising Online Media on COVID-19 during the Early Months of the Pandemic. *Journal of Quantitative Description: Digital Media*. Forthcoming.