Charles Alba

314-861-7674 | alba@wustl.edu | sites.wustl.edu/alba | Google Scholar | cja5553 | charles-alba-24bb63186

Education

PhD in Computational & Data Sciences — *St Louis City, MO, USA*
Washington University in St Louis — 2022 - present
- GPA: 4.0/4.0, Danforth Scholar

MS in Behavioral & Data Science (Distinction) — *Coventry, England, UK*
University of Warwick — 2021 - 2022
- Thesis: "The role of default-displayed reviews in anchoring online reviewers: text evidence from STEAM video-games"

BS in Data Science — *University Park, PA, USA*
The Pennsylvania State University — 2018 – 2021
- Minor in Statistics, GPA: 3.83/4.0, Mu Sigma Rho Honors, 7× Deans List, Remote Innovation Grant recipient

Skills

Technical skills Machine learning & AI | deep learning | natural language processing (NLP), text mining & large language models (LLMs) spatial mapping & analytics | statistical models & analysis | Data wrangling | web scrapping & using APIs

Coding skills Python (including `sk-learn`, `pandas` & `Pytorch`) | R (including `tidyverse`) | SQL | NoSQL languages | QGIS | LaTeX

Languages English (Native) | Chinese (proficient)

Experience

Graduate Research Assistant — *St Louis City, MO, USA*
Division of Computational & Data Science, Washington University in St Louis — Aug 2022 - present
- Research areas include:
  - Applying and tuning state-of-the-art *large language models (LLMs)* to build predictive models across the social science & healthcare sectors.
  - Applying statistical tools to novel data sources (eg text data) to provide inferential insights in the social science & healthcare domains.
- Selected research projects:
  - Applying *large language models (LLMs)* on clinical notes to predict complications from patients undergoing perioperative care.
    - We used ∼ 85k clinical notes from the BJC Healthcare system to build predictive LLMs aimed at predicting surgical complications and outcomes from patients undergoing surgery.
    - We employed state-of-the-art pre-trained, clinically-oriented LLMs, such as `BioGPT`, `ClinicalBERT`, and `bioClinicalBERT`. Our experiments also involved novel fine-tuning approaches, including semi-supervision and the development of foundational models through multi-task learning (MTL), utilizing `Pytorch`.
    - These approaches have led to performance enhancements over traditional foundational models and fine-tuning methods. Notably, our models have achieved improvements in predictive outcomes, with up to a 3.6% increase in AUROC and 2.6% in AUPRC.
    - Status: Preparing a manuscript for submission to NEMJ AI. Code available at github.com/cja5553/LLMs_in_medicine.
  - Building LLMs to understand emotions and sentiments concerning food-assistance policy in the United States.
    - Most food assistance policies, such as SNAP, are evaluated through questionnaires, making them costly and labor-intensive. Additionally, feedback concerning this topic is rare, making it difficult for policymakers to discern the popularity of specific policy changes toward SNAP.
    - To address this, we plan to collect ∼50k relevant articles across major news outlets spanning 5 years. We intend to classify news headlines based on sentiment (i.e., positive or negative), as well as emotions (i.e., happy, sad, fearful, or angry). This allows us to understand the favorability of specific policy changes concerning SNAP.
    - We intend to experiment with parameter-efficient fine-tuning methods, such as LoRa, IA3, prefix-tuning, and P-tuning, on top of traditional fine-tuning.
    - Status: Currently liaising with external vendors to retrieve the dataset (i.e., newspaper articles).
  - Using *mobile phone data* to assess socio-economic disparities in unhealthy food consumption during COVID-19
    - Cleaned and aggregated county-level longitudinal data on visits to unhealthy food outlets from ∼80k points-of-interests using `Python`.
    - Merged this data with the New York Times' COVID-19 incidence data and socio-economic data from the American Community Survey.
    - Employed the logit fixed-effects model in `R` to analyze the impact of COVID-19 on changes in socio-economic disparities associated with unhealthy food reliance.
    - Status: Manuscript published at Health Data Science journal. Code available at github.com/cja5553/mobile-phone-data-to-assess-unhealthy-food-reliance. Manuscript available at spj.science.org/doi/abs/10.34133/hds.0101.

Graduate Student Assistant — *Coventry, England, UK*
Dept of Psychology, University of Warwick — Mar 2022 - Aug 2022
- Project: Applying *text-mining* and *NLP* tools to examine attention allocation and cognitive biases in online reviews of video games
  - We hypothesized that more salient and higher-ranked reviews have a more profound influence on users who write new reviews.
  - After scrapping over ∼1.1 million video game reviews from STEAM, I reverse-engineered STEAM's review sorting algorithm (with an accuracy of ∼0.7 kendalls τ coefficient) to identify reviews that would have been shown to each user when they were writing their own review.
  - I then employed FastText embeddings and cosine-similarity to understand the similarity between reviews written by new reviewers with the reviews that would have been displayed to them whilst they were writing their reviews with `Python`.
  - ANOVA results revealed the display of salient and top-ordered reviews significantly influenced subsequent review content.
  - Status: Manuscript under revision at Decisions journal. Code available at github.com/cja5553/attention-driven-imitation-in-consumer-reviews,

Undergraduate Research Assistant — *University Park, PA, USA*
Dept of Recreation, Parks & Tourism Management, Penn State University — May 2021 – Aug 2021
- Project: Using *mobile phone data* to assess the impact of COVID-19 on inequity in recreational tourism.
  - Analyzed COVID-19's impact on inequity towards national park visits using mobile phone data.
  - I cleaned, filtered and selected data from over ∼40 million points-of-interest (POI) using spatial tools in `R`, `Python` and `QGIS`.
  - Integrated US census data for demographic insights and employed the 'gravity-model' panel data analysis with `R`.
  - Observed increased in visits from communities located less than 452km of a National Park.
  - However, a significant decreased visitations were witnessed from non-white and Native American communities, especially is they are located more than 317km and 482km from a national park, respectively.
  - Article and code available at Scientific Reports: doi.org/10.1038/s41598-022-16330-z