ERIC BOXER (He/Him)

ecboxer.github.io

Email : eboxer@usc.edu
Mobile : (347) 525-3516

Research Statement

I am interested in making powerful machine learning algorithms (such as large language models) more explainable and investigating how AI can learn in a more human-understandable manner. In my professional experience, I worked on problems in event detection, information extraction, and health assessment. During that time, questions about reliability and understandability cropped up repeatedly and often limited the impact that could be made with even the most performant machine learning solutions. I believe that the AI field is currently overcoming "technophobia" by sheer performance and novelty but that may eventually work against us if we expect regular folks to incorporate agents they don't understand into their day-to-day lives.

Education

- **University of Southern California** — Los Angeles, CA
  PhD in Computer Science (Advised by Jay Pujara) — *Aug 2023 – Present*
  Courses: Advanced Analysis of Algorithms, Ethics in NLP
- **Columbia University** — New York, NY
  MS in Data Science; GPA: 3.8 / 4.0 — *Aug 2018 – Dec 2019*
  Selected Courses: Causal Inference, Network Analysis, Algorithms, Statistical Inference and Modeling
- **SUNY Stony Brook** — Stony Brook, NY
  BS in Mathematics and Economics, Minor in Philosophy; GPA: 3.9 / 4.0; Phi Beta Kappa — *Aug 2013 – Dec 2016*
  Selected Courses: Symbolic Logic, Differential Equations, Topology and Geometry, Game Theory

Publications

- N. Le Vine, **E. Boxer**, M. Dinani, P. Tortora, and S. Das. Identifying early warning signals from news using network community detection. In *AAAI Conference on Innovative Applications of Artificial Intelligence*, 2022.

Reviewing Activities

- **CIKM** *Long/Full Papers and Short Papers* — 2023
- **CIKM** *Applied Research Track* — 2022

Presentations

- "Early Warning Signals at Swiss Re" Big Data Minds EU — 2022

Professional Experience

- **USC Information Sciences Institute** *Graduate Research Assistant* — August 2023 – Present

  Modeling Information Pathways: Developing an approach to modeling the flow of news among different publications and social media platforms using temporal knowledge graphs (TKGs). In particular, modeling reporting on the conflict between Russia and Ukraine. Incorporated warm-starting from language models into existing TKG representation methods and experimenting with regimes for best balancing the information gained from the language models and generalizability. I am also interested in using this setup to research explainable link prediction and approaching news as interconnected/interdependent data rather than discrete events.

- **Swiss Re** *Data Scientist* — Jan 2020 – July 2023

  Early Warning Signals: Developed a Python package to extract "signals" from a corpus of news articles. Viewing the corpus as a network (with keywords as nodes and co-occurrence of keywords in an article forming edges), we extract communities from the network and present important articles for internal users in a dashboard. Article importance is determined by network-centric metrics, including their "novelty" as measured by a link prediction model. Presented the method at AAAI and internal/external industry events to technical and non-technical audiences.

  Sentiment: Developed a repeatable aspect-based sentiment analysis (ABSA) approach for topics such as COVID-19 medical developments and financial news. By either gathering a set of articles with "noisy labels" for pre-training or starting with an open-source pre-trained domain-relevant model, we finetuned ABSA models for our topic of interest. Model output was presented in a dashboard to medical experts and analysts to curate their news consumption.

  Excel Extraction: Developed models to extract variable-format tabular data from Excel files, containing information about properties in insurance submissions. We frame this as a sequence of NLP tasks (named entity recognition and text

categorization) and use 5 such independent models feeding into two tree-based metamodels which provide final predictions. Also, trained models to normalize free-text values (i.e. multi-class classification) into domain-compliant categories. The full pipeline (extraction plus normalization) is in production and can be used to gradually replace the current manual entry of Excel documents or as part of a human-in-the-loop process.

Alternative Data Sources: Trained interpretable models to predict underwriting outcomes using traditional life insurance data (e.g. application responses, prescription histories, driving records, past insurance history) and new 3rd-party data sources (e.g. electronic health records, billing records, clinical labs), to assess the ability of new data sources to replace traditional sources, for multiple clients. Built a simple UI with Dash to collect labels for the project.

Accelerated Underwriting: Trained an interpretable model to reduce referrals by 30%, for a mid-sized US life insurer, thereby reducing costs. Delivered applicant cohorts (created with unsupervised self-organizing maps) and insights about the client's underwriting process that were incorporated into their underwriting rules.

Culture: Organized monthly meetups on data science-related projects for Swiss Re colleagues. Reached out to potential speakers throughout the company and externally, hosted the meetups, and wrote summary blog posts for internal engagement. Led intern- and early career-recruiting efforts for data science teams in the Americas and acted as a "buddy" providing support to about ten new joiners over my time with the company.

- **Swiss Re** *Data Science Intern* May 2019 – Dec 2019

  Early Warning Signals: Developed a novel outlier detection method for news articles, to identify potentially critical information for decision makers. Viewing the corpus as a network, we are able to identify novel articles as those which form links between otherwise distant nodes. Also, extended the node2vec method of computing node embeddings to include domain knowledge (by warm-starting node embeddings) and early stopping (by computing link prediction scores on a hold-out validation network).

  Property & Casualty Solutions: With financial and geospatial data from the client's portfolio, explored drivers of named-event loss (i.e. caused by significant events such as fires or floods) and attritional loss. Integrated in-house natural catastrophe models and publicly available fire incident data to create and validate policy- and broker-level risk scores. Recommendations were presented to the client, who expressed interest in continued engagement with the analytics team and reinsurance renewal with Swiss Re.

- **Frac.tl** *Data Visualization Developer* March 2019 – June 2019

  Created visualizations for articles that drove search traffic for the firm's clients. Scraped websites for data that was not freely available in tabular format such as sports team attendance and player wages. Created interactive and static data visualizations in RShiny and ggplot2. For example, an interactive map of MLB ticket prices over the period 1950 to 2017, with icon size as a function of ticket prices and icons placed according to team location at the time.

Programming Skills

- **Python**: PyTorch, Keras, Scikit-Learn, spaCy, pandas, numpy, scipy, dash
- AWS, Azure, D3, Elasticsearch, Emacs, Git, Javascript, Jupyter, LaTeX, Linux, PySpark, R, SQL, VSCode