Advancing Salary Transparency Through NLP: A Comparative Study of Two Approaches for Predicting Salaries from Job Descriptions

Team Member (Solo Project): Zebang Li (Duke NetID: zl411)

Problem Statement

Despite growing demands for wage transparency, only a handful of states—including California, Washington, and New York—currently mandate salary range disclosure in job postings. This leaves job seekers in most states navigating career decisions with limited compensation information. While legislative changes progress slowly, machine learning offers an immediate opportunity to bridge this information gap.

This project aims to develop and compare two NLP approaches for predicting salaries from job descriptions. While existing solutions often rely on structured data like job categories, titles and years of experience, this project explores the potential of extracting salary-relevant information directly from unstructured job description text.

Approach

The project will utilize the job posting data on LinkedIn, which contains:

- Detailed job descriptions
- Associated salary ranges (min, max)
- Additional metadata (location, industry, etc.)

I will implement and compare two distinct approaches:

Approach 1: Fine-tuned BERT Model

- Base model: A pre-trained transformer model (BERT)
- I will fine-tune a pre-trained transformer model, specifically BERT, and implement a regression head to predict continuous salary values.
- The project will also experiment with different BERT variants (e.g., BERT, RoBERTa, DistilBERT) and hyperparameters to optimize performance.

Approach 2: Custom Neural Network with Pre-trained Embeddings

- Word embeddings: Word2Vec or GloVe pre-trained embeddings
- Model architecture: A custom neural network with LSTM layers. Add dense layers to transform the extracted features into the final salary prediction.