CIS 6930: Trustworthy Machine Learning
Project Proposal: Decrypting job title classification model

Tanvi Jain
(Point of Contact)
tjain@ufl.edu

Nimish Bajaj
nimishbajaj@ufl.edu

September 21, 2021

1 Introduction

Determine whether explainable/interpretable ML techniques like LIME and others can provide useful insights into the cause of the unfairness.

A model's accuracy is not always enough to state whether it will perform well in the wild because the accuracy highly depends on the data that was used to train and test the model. Explaining the results of a model, and identifying what leads to a particular classification can help provide insights into the model. Once you add interpretability to the model, its result can be easily understood by a domain expert can be verified for correctness. This greatly enhances the trustworthiness of the model.

We are proposing to build a text classification model and explaining the results using LIME, and validate if LIME can provide stable and meaningful features.

2 Background and Related Work

Model interpretation techniques have been in place for the last couple of years, but they are not always able to provide stable inferences for a given problem. Work done on inference healthcare model[3], Perturbing Inputs for Fragile Interpretations in Deep Natural Language Processing [5] and many more have a common objective, i.e. to have robust interpretations for determining the trustworthiness of the model.

For our project, we are planning to decrypt the Job classification model on Resume corpus. Our work will include categorizing resumes using CNN and RNN. We are going to train Neural Network and use the inference technique 'LIME'. Since the data contains technical words which are sometimes outside of the English vocabulary, we are going to tune word embedding for the problem domain.

Existing work has explored a CNN for training a classification model, as part of the project we will be using an RNN to learn the temporal features. The data contains technical words which are sometimes outside of the English vocabulary, we are going to tune word embedding for the problem domain by utilizing the training dataset and other datasets from the same domain.

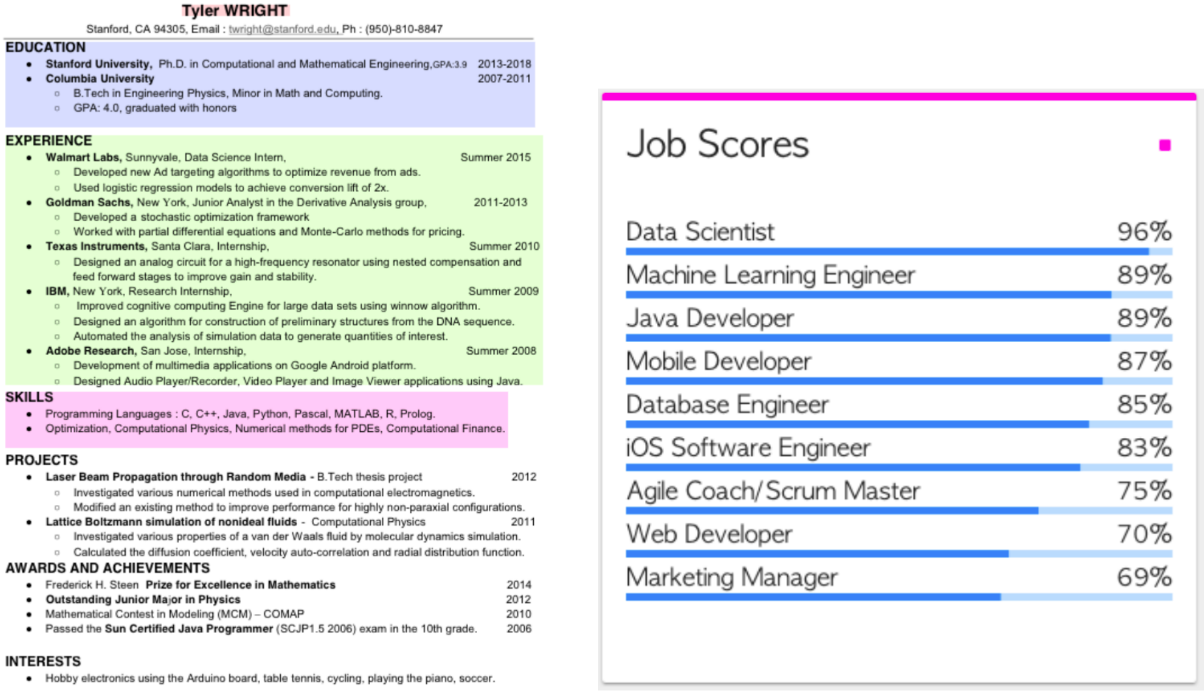
3 Proposed Approach & Plan

Following is the approach we are going to take:

1. Fetch data and take samples as per the distribution of data for train, test, and validation
2. Clean and vectorize the data for Neural Network (NN) approach
3. Build a NN architecture to learn classification labels
4. Train, test, and validate the neural network
5. Use LIME for model inferencing

6. Manually validate LIME output

For the test and experimentation, we are currently foreseeing that the local machine would be sufficient for carrying out the work. If needed we will use the hyper-gator resources that are availed to us as part of this project.



Tyler WRIGHT

Stanford, CA 94305, Email : twright@stanford.edu, Ph : (950)-810-8847

EDUCATION

- **Stanford University,** Ph.D. in Computational and Mathematical Engineering,GPA:3.9 2013-2018
- **Columbia University** 2007-2011
  - B.Tech in Engineering Physics, Minor in Math and Computing.
  - GPA: 4.0, graduated with honors

EXPERIENCE

- **Walmart Labs,** Sunnyvale, Data Science Intern, Summer 2015
  - Developed new Ad targeting algorithms to optimize revenue from ads.
  - Used logistic regression models to achieve conversion lift of 2x.
- **Goldman Sachs,** New York, Junior Analyst in the Derivative Analysis group, 2011-2013
  - Developed a stochastic optimization framework
  - Worked with partial differential equations and Monte-Carlo methods for pricing.
- **Texas Instruments,** Santa Clara, Internship, Summer 2010
  - Designed an analog circuit for a high-frequency resonator using nested compensation and feed forward stages to improve gain and stability.
- **IBM,** New York, Research Internship, Summer 2009
  - Improved cognitive computing Engine for large data sets using winnow algorithm.
  - Designed an algorithm for construction of preliminary structures from the DNA sequence.
  - Automated the analysis of simulation data to generate quantities of interest.
- **Adobe Research,** San Jose, Internship, Summer 2008
  - Development of multimedia applications on Google Android platform.
  - Designed Audio Player/Recorder, Video Player and Image Viewer applications using Java.

SKILLS

- Programming Languages : C, C++, Java, Python, Pascal, MATLAB, R, Prolog.
- Optimization, Computational Physics, Numerical methods for PDEs, Computational Finance.

PROJECTS

- **Laser Beam Propagation through Random Media -** B.Tech thesis project 2012
  - Investigated various numerical methods used in computational electromagnetics.
  - Modified an existing method to improve performance for highly non-paraxial configurations.
- **Lattice Boltzmann simulation of nonideal fluids** - Computational Physics 2011
  - Investigated various properties of a van der Waals fluid by molecular dynamics simulation.
  - Calculated the diffusion coefficient, velocity auto-correlation and radial distribution function.

AWARDS AND ACHIEVEMENTS

- Frederick H. Steen **Prize for Excellence in Mathematics** 2014
- **Outstanding Junior Major in Physics** 2012
- Mathematical Contest in Modeling (MCM) – COMAP 2010
- Passed the **Sun Certified Java Programmer** (SCJP1.5 2006) exam in the 10th grade. 2006

INTERESTS

- Hobby electronics using the Arduino board, table tennis, cycling, playing the piano, soccer.

(a) Example Input

Job Scores

Data Scientist	96%
Machine Learning Engineer	89%
Java Developer	89%
Mobile Developer	87%
Database Engineer	85%
iOS Software Engineer	83%
Agile Coach/Scrum Master	75%
Web Developer	70%
Marketing Manager	69%

(b) Model prediction

Figure 1: Model output description [3]

4 Timeline

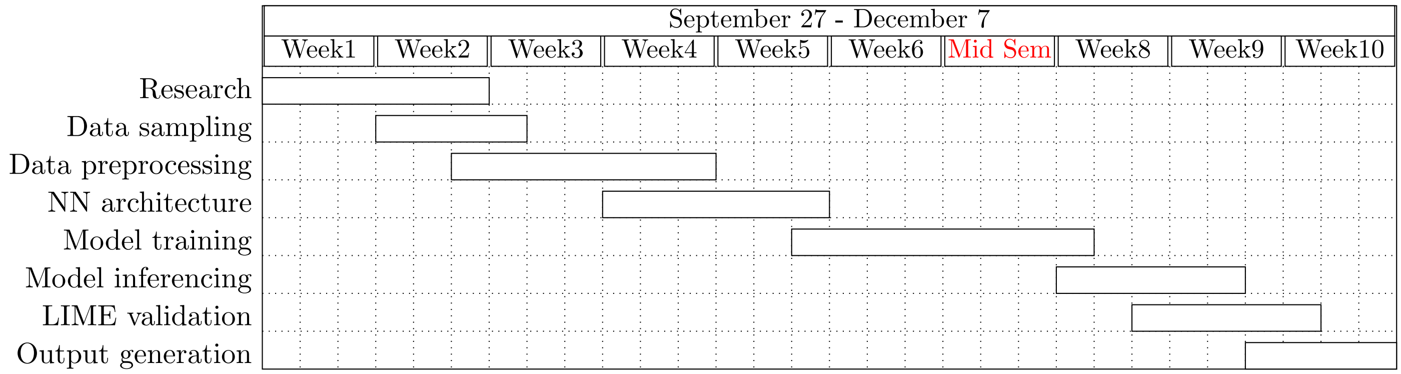


Figure 2: Gantt Chart

References

[1] Damien Garreau and Ulrike von Luxburg. Explaining the explainer: A first theoretical analysis of lime. 2020.

[2] Kameni Florentin Flambeau Jiechieu and Norbert Tsopze. Skills prediction based on multi-label resume classification using cnn with model predictions explanation. *Neural Computing and Applications*, 33(10):5069–5087, May 2021.

[3] Khansa Rasheed, Adnan Qayyum, Mohammed Ghaly, Ala Al-Fuqaha, Adeel Razi, and Junaid Qadir. Explainable, Trustworthy, and Ethical Machine Learning for Healthcare: A Survey. 4 2021.

[4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should i trust you?”: Explaining the predictions of any classifier. 2016.

[5] Sanchit Sinha, Hanjie Chen, Arshdeep Sekhon, Yangfeng Ji, and Yanjun Qi. Perturbing inputs for fragile interpretations in deep natural language processing, 2021.

[6] Tim Zimmermann, Leo Kotschenreuther, and Karsten Schmidt. Data-driven hr - resume analysis based on natural language processing and machine learning, 2016.