Scam Detection using NLP

DS 4002 – Fall 2023 - Professor Alonzi – Created by Peter Layne
Due: 2 weeks after start
Submission format: Upload link to github repo to canvas

Individual Assignment

General Description: Submit a github repo containing your work plan, ALL code attempts, and results

Preparatory Assignments – None, some background in NLP helpful

Why am I doing this? This is an opportunity

- Course Learning Objective: Gain experience using NLP models to build a model that will have real world applicability

What am I going to do? In this assignment you will utilize available data to build a machine learning model to catch fake job posting. You will need to draw on your experience with machine learning techniques and utilize the internet to learn more about NLP if you do not have prior experience. You will also need to draw on your knowledge of ML evaluation to evaluate your model and explain where it's gaps may be.
Deliverables include:

- Github repository containing your code, any additional data used, links to any outside sources consulted, figures demonstrating that you have evaluated your model
- A brief, 1 page write-up of the project addressing what you have learned throughout the project and how you plan to apply this to future projects

All of this will be submitted electronically via canvas

Tips for success:

- Get creative! There is no right model to create, if you can find a way to incorporate more data, do it!
- Don't be overly ambitious: if you have no experience in NLP, start slow. It is unlikely that you are going to build a model that could be used as an enterprise solution!
- Embrace the challenge: especially if you have never used NLP models, this type of model can be hard. Utilize the resources available to you (Professors, TA's, Peers, Stack Overflow) and take pride in grinding this out
-

How will I know I have Succeeded? You will meet expectations when you have followed the rubric below:

Spec Category	Spec Details

Formatting	• Repository – A github repo (and cloud storage folder if necessary) containing all materials
	  ○ Submit a link to the repo
	  ○ Everything is contained in the repo or linked to it if appropriate
	  ○ Contents
	  ▪ Additional Data
	  ▪ Code
	  ▪ Sources of Code consulted
	  ▪ A folder of figures (should include your evaluation metrics)
	  ○ Use pdf format when possible
	  ○ For code and data products use the appropriate format for whatever it is
	• One Page Writeup- Submitted to canvas as a PDF with name First_Last_writeup

Repository	• Goal: This repository should represent the core of your project. • Code: Please include all data cleaning, partitioning and preprocessing. Include all code for building the model and evaluating its performance. • Source Code: please include citations or links for all articles or outside sources used to help build the model • Please include screenshots of outputs evaluating model performance, this way if we have issues replicating your code we can still see these figures.
Writeup	• Goal: To reflect on the process of building and creating the model and how you could improve in the future • Include an executive summary that goes through the main points of the reflection • Include a paragraph reflecting on the biggest challenge of the project • Include a paragraph reflecting on what you would change if you could do the project again

Acknowledgements: Special thanks to Jess Taggart from UVA CTE for coaching on making this rubric. This structure is pulled direction from Streifer & Palmer (2020).