adept ID

Data Science - Technical Exercise
November 2023

Background

A common task we encounter in our work is the classification of job titles and postings into occupational categories (O*NET)[1]. This categorization allows us to infer and append labor market data collected by the Bureau of Labor Statistics (BLS) such as expected wages and educational requirements.

In the past, many providers of labor market data would perform this classification task with complicated, rules-based systems. However, recent advances in language models allow us to take advantage of the capacity of these models to produce semantic features or "embeddings" making this task amenable to machine learning approaches.

Problem Statement

In this take-home, we provide two datasets of job postings and their associated O*NET classifications. The first dataset contains classified postings for training and model development. The second dataset contains an identically structured dataset for testing.

In this exercise, you will use the provided data to construct a model which ingests job postings and returns the top N most likely occupational categories (O*NETs) for each job posting. This model should leverage a pre-trained language model of your choosing. Many such models can be found on the HuggingFace platform. While we don't have specific requirements regarding the exact model (you can choose any you like), this model may be helpful for this assignment.

Your implementation should include two primary components:

1. A **predict()** function or method that takes in a job posting and returns the top N most likely O*NETs.
2. A set of metrics and associated charts that describe the effectiveness of your classification model.

Along with your code, please put together a brief memo describing your approach and how you would extend this approach to improve the classification accuracy were this an actual model development project rather than a time-boxed, take-home exercise. We're primarily interested in understanding the logic behind your modeling approach so please emphasize this in your deliverable.

[1] The full taxonomy can be found here.

Please note: We understand the limitations of this time-boxed exercise. The overall accuracy of your models is *not* of particular importance to this exercise so *please don't spend too much time optimizing for model accuracy*. The goal is to evaluate your comfort level with thinking critically about system design and associated metrics for measurement.

I hope that you find this exercise interesting! Best of luck!

Data Provided

1. **train_data.csv** - This file contains sample parsed job postings and their associated O*NET classifications for model development. This data has the following fields:
    a. ID - A unique identifier for the posting
    b. POSTED - The date the job was posted
    c. TITLE_RAW - The parsed raw job title from the posting
    d. BODY - The body text of the posting
    e. ONET_NAME - The classified occupation title
    f. ONET - The classified occupation ID
2. **test_data.csv** - This file contains an identical format to train_data.csv and should be used for model evaluation.

Deliverables

1. Code - Github repository containing the code you wrote to complete the assignment.
2. Memo - Describe your approach, key assumptions, metrics and conclusions. Importantly, describe how you would expand this analysis were it a complete project with a larger timeline.

What this tests

1. Can you take an abstract idea and implement it in code?
2. Can you construct metrics to evaluate the performance of your models?
3. Can you identify, reason through and address constraints from real world problems?
4. Can you communicate findings clearly?