PROJECT PROPOSAL

Udacity AWS Machine Learning Engineer Nanodegree Capstone

BY SATYAM CHATROLA

Domain Background:

Essays represent a person's critical thinking. It provides a glimpse into a person's mind. This serves as an approach to cultivate thoughts and ideas, and also a way to evaluate a person. This method is used by academia and educational institutions. Essay evaluation has been performed manually by humans, which is subject to many factors and could lead to inconsistency.

Problem Statement:

We would be given essays that are rated by human evaluators. And our task would be to develop an AI model that can rate it as accurate as possible compared to a human evaluator.

Solution Statement:

We would experiment with multiple open-source ML models such as LSTMs, Transformers and fine-tuned LLMs like GPT-2. The models could be pre-trained and fine-tuned, or they could be trained from scratch. It is a classification problem in Natural Language Processing (NLP) domain.

Datasets and Inputs:

The Automated Student Assessment Prize (ASAP) dataset, consisting of 24,000 graded student essays, will serve as the primary data source. Essays span a variety of prompts and subjects, providing a robust basis for evaluating the generalizability of our models. Each essay was pre-processed to normalize text, remove non-alphabetic characters, and tokenize sentences.

Benchmark Model:

The Benchmark Model is the Ordinal Class Classifier [Frank and hall, 2001] used in Mathias, S. A.; and Bhattacharyya, P. 2018. ASAP++: Enriching the ASAP Automated Essay Grading Dataset with Essay Attribute Scores. They were able to get maximum Kappa score of 0.76. This is our benchmark to compare our results.

Evaluation Metrics:

We evaluate each of the annotators using Cohen’s Kappa, with quadratic weights - i.e. Quadratic Weighted Kappa (QWK) (Cohen, 1968). We chose this as the evaluation metric (as compared to accuracy and weighted F-Score) because of the following reasons:

1. Unlike accuracy and F-Score, Kappa takes into account random agreement. For example, a majority class classification will result in a Kappa of 0, while accuracy and F-Score will be the percentage of the majority class in the test set.

2. Weighted Kappa takes into account the distance between the actual score and the reported score. Quadratic weights reward matches and penalize mismatches more than linear weights.

Project Design:

We will create a webapp. It will be a form where the users provide input (their essay, and the evaluator AI model [LSTM, BERT, GPT-2] that will evaluate the essay.) We would be using a FastAPI backend server that will serve requests.