Project Report

Detecting AI Generated Text

The Problem

In recent years, large language models (LLMs) have become increasingly sophisticated, capable of generating text that is difficult to distinguish from human-written text. This project aims to develop a machine learning model that can accurately detect whether an essay was written by a student or an LLM. The competition dataset comprises a mix of student-written essays and essays generated by a variety of LLMs.

The Solution Overview

The objective of was to develop and evaluate two different approaches for detecting AI-generated text. The two approaches implemented are :

a) Ensemble learning, combining multiple models.
b) Utilizing pre-trained BERT (Bidirectional Encoder Representations from Transformers).

Dataset used:

The given dataset comprises of both human-written and AI-generated samples. Another dataset was concatenated with the original one to create a wider set of training data. The 'prompt_names' were replaced with integers as 'prompt_id'.

a) The prompt_id column denotes which prompt was used to generate the data
b) Text column has the text content of the essays
c) Generated labels indicated whether the text is AI-generated(1) or human-written(0).

Training

Approach – 1: Ensemble Learning:

a) The text from the data was tokenised and vectorised first, using the word.tokenise from NLTK and the Count Vectorizer from scikit-learn, respectively.
b) Ensemble method used: Voting Classifier from scikit-learn.
c) Individual models:
    - Model 1: Multinomial Naïve-Bayes classifier
    - Model 2: Stochastic Gradient Descent Classifier
    - Model 3: Light Gradient Boosting Machine
d) Weighted combination to leverage strengths of individual models.
e) The data was split into train and test data and was fitted into the ensemble model.

Approach – 2: Using BERT:

a. The pre-trained BERT model was employed, specifically the "bert-en-uncased-l-12-h-128-a-2" version, obtained from TensorFlow Hub.
b. A model was then built – a combination of the pre-trained BERT model for contextual embeddings and additional dense layers for task-specific feature learning. The BERT model was also fine-tuned as required.
c. The same dataset was used, after splitting into train and test, and the model was trained on it.

The Evaluation

Ensemble Approach:

- Evaluated on metric – accuracy on the test set.
- Accuracy was measured to be 0.9932972972972973

BERT-based Approach:

- Evaluated on metric – accuracy on the test set.
- Accuracy was measured to be 0. 0.9878910183906555

Comparison:

Of the two, the BERT model-based approach gave poorer accuracy, and required a lot more time and space. Thus, the ensemble method is the better choice here.

Challenges:

a. Handling imbalanced datasets.
b. Fine-tuning BERT efficiently due to computational requirements.

Conclusion:

This project successfully addressed the challenge of detecting AI-generated text using two distinct approaches - ensemble learning and BERT-based text models. The comparative analysis provides insights into the strengths and limitations of each approach, paving the way for further advancements in this field.

Submitted by:
Shoilayee Chaudhuri
21112102
Chemical Engg – 3y