Hackathon Solution Documentation

Solver: Manoj Kumar

This document outlines the approach taken to solve the given hackathon challenge, which involved distinguishing between essays written by students and those generated by a Language Model (LLM).

1. Data Exploration:

The data provided includes 'train.csv', 'test.csv', and 'train_prompts.csv'. Each file was examined to understand its structure and contents.

2. Data Preprocessing:

The training and test datasets were merged with the prompts data on 'prompt_id' to include details such as prompt name, instructions, and source text alongside the essays.

3. Feature Engineering:

Key features extracted from the essay texts include text length, word count, and average word length. These features aim to capture the essence of the text that might help in distinguishing between human and AI writings.

4. Model Building and Evaluation:

A RandomForestClassifier was used for this task. The model was trained on the training set and evaluated using the ROC AUC metric, achieving a score of approximately 0.997 on the validation set.

5. Submission Preparation:

Predictions were made on the test set, and a submission file was prepared following the provided example. The submission file contains the essay 'id' and the predicted probability of the essay being generated by an AI.

The approach emphasized simplicity and effectiveness, leveraging basic text features to achieve high predictive

accuracy.

This solution was crafted by Manoj Kumar, demonstrating a methodical approach to tackling the challenge posed by the hackathon.