A Model for Distinguishing Essays Generated by Large Language Models from Student-Written Essays

In recent years, the proliferation of large language models (LLMs) has raised significant concerns about their potential misuse, particularly in educational settings. This project, developed by me, aims to address these concerns by developing a model capable of discerning essays generated by LLMs from those authentically authored by middle and high school students. At the heart of this initiative lies the ambition to mitigate issues related to plagiarism and its potential adverse impact on students' skill development.

Dataset:

The foundation of this project is a meticulously curated dataset comprising essays labeled as either generated by LLMs or composed by students. This annotated corpus serves as the linchpin for supervised learning, allowing the model to learn and generalize patterns inherent in LLM-generated and student-written content.

Text Processing:

To extract meaningful features from the textual data, a TF-IDF vectorizer is employed. This widely used technique transforms the essays into numerical vectors, capturing the significance of words relative to a larger collection of documents. The TF-IDF vectorizer acts as a crucial preprocessing step, laying the groundwork for subsequent model training.

Model Selection:

For the task of distinguishing between LLM-generated and student-written essays, a Random Forest Classifier is chosen. This ensemble learning method constructs multiple decision trees during training, culminating in a robust model capable of discerning the nuanced patterns that differentiate the two categories.

Training and Evaluation:

The model undergoes rigorous training using the processed dataset, adjusting its parameters to learn the intricate characteristics of LLM-generated and student-written essays. Evaluation is conducted on a validation set, employing the ROC AUC scoring metric. This metric provides a comprehensive assessment of the model's performance in distinguishing between the two categories, offering insights into its efficacy and reliability.

Importance of the Project:

This project underscores the critical need to advance LLM text detection, particularly in educational contexts. By addressing concerns related to plagiarism and its potential impact on student skill development, the initiative aims to contribute significantly to the preservation of academic integrity in an era where technological advancements pose both opportunities and challenges.

Tool for Educational Contexts:

The resulting model is envisioned as a pivotal tool for educational contexts. Whether integrated into educational platforms or utilized by teachers and institutions, this model stands as a safeguard against the inadvertent or deliberate use of LLMs in students' work. It not only ensures the authenticity of student contributions but also fosters an environment that upholds the principles of fair learning and academic honesty.

In conclusion, this collaborative effort represents a multifaceted approach to addressing the challenges posed by LLMs in educational settings. By leveraging state-of-the-art techniques in text processing and machine learning, the project aims to provide a tangible solution that safeguards academic integrity and empowers educators and institutions in navigating the evolving landscape of education in the digital age.