BT5153

Group Project Proposal

Detecting AI-Generated Text for Evaluation on Academic Integrity

Fang Ziwei (A0130515W)

Janani Jayaraman (A0280567W)

Mao Yuxin (A0129810J)

Shi Yiran (A0262832E)

Zhang Jiajia (A0130345U)

Intent

With the advancement of large language models like Chat GPT and Gemini, it has become easier for students and researchers to generate text that mimics human writing on any possible topic under the sun, leading to potential instances of plagiarism and academic dishonesty. Traditional methods of plagiarism detection may struggle to distinguish between human and AI-generated text, creating a need for specialized tools to address this issue. By developing a system capable of detecting AI-generated text, we could enhance the credibility and integrity of academic work.

Our target customers are educational institutions looking for a reliable tool that helps identify instances of AI-generated text in academic submissions. Institutions committed to upholding academic integrity can leverage this technology to demonstrate their fair assessment of students' work, dedication to scholarly rigor and originality. This will help enhance their reputation and credibility within the academic community and among stakeholders who value integrity in research and education. By developing an AI-generated text detector, we contribute to the preservation of academic standards, the cultivation of a culture of honesty and integrity in academia. Moreover, we create an environment that fosters genuine knowledge gain and encourages students to produce original work that contributes meaningfully to their fields and in turn promotes innovation and originality.

Desired Outcome

Performance of our product will be measured by the proportion of plagiarism cases correctly detected. Success of our product lies in surpassing established benchmarks, notably those set by commonly used plagiarism-checking tools in educators worldwide like Copyleaks and GPTZero. Our success metrics revolve around the increased percentage of AI-generated instances identified by our model. By comparing the performance of our product against existing plagiarism detectors, we seek to offer a better or equally accurate prediction model, yet a more cost-efficient and lightweight solution for academic institutions.

Deliverables

Data Collection and Preparation

Assemble and refine a comprehensive dataset of both AI-generated and human-written texts for training and evaluating the detection model. This involves sourcing data that covers a wide range of vocabulary, styles, and complexity levels (for example, augmented data from Kaggle competition, Hugging Face, and student essays competitions). This dataset will be divided into train, validate and test subsets to prepare for effective model training and evaluation. Afterwards, the train dataset will then undergo preprocessing to normalize, tokenize, and possibly vectorize (embedding) the texts.

Model Development and Evaluation

We plan to develop a baseline model to facilitate evaluations and gauge the improvements. Combining TF-IDF vectorizer with a logistic regression classifier could be used as an effective baseline in distinguishing AI-generated texts.

More advanced models such as DistillBERT, XLNet and Transformer-XL will be considered as candidates to explore a broad spectrum of detection strategies. The model that best balances accuracy with

computational efficiency will be selected for further optimization, with a special emphasis on improving its specificity without significantly compromising its detection capabilities.

The selected model will undergo a thorough evaluation to ensure its performance and reliability. This evaluation will leverage key metrics such as accuracy, precision, recall, and the F1 score, with a special emphasis on minimizing false positives. For example, The **AUC-ROC** curve and confusion matrix analysis help in assessing the model's discriminative ability and identifying conditions that may lead to false identifications. A **cost-benefit analysis** quantifies the impact of false positives versus true positives, guiding model adjustments to minimize wrongful identifications. Additionally, robustness and sensitivity analysis will test the model's performance across various conditions to ensure stable minimization of false positives.

Enhancement through Data Augmentation

After the initial model evaluation, we plan to augment the dataset by using AI (e.g. ChatGPT 4.0 and Copilot) to paraphrase human-written content gathered from textbooks across different academic disciplines, labeling these paraphrased versions as AI-generated and keep the discipline as a slicer for evaluation. By training the model on similar content across both genuine and AI-manipulated texts, we intend to improve its ability to discern the difference between the two categories. Meanwhile the discipline slicer would be providing assessment for its effectiveness and applicability in diverse academic settings.

Model Deployment and Test

Develop and launch a standalone web application with simple UI to enable users to input text or upload files for analysis. The application should utilize the trained model to assess whether the content is AI-generated, providing results directly on the UI along with a probability score.

The model application will undergo comprehensive testing using the test dataset, including both functional and performance assessment, to ensure its reliability and efficiency. This deployment aims to make AI-generated text detection readily available for immediate content evaluation.

Constraints

One of the primary challenges in plagiarism detection could be the difficulty in identifying instances where contents generated by AI have been rephrased by humans. Human paraphrasing AI-generated text would have altered the language pattern that the ML model is capturing, hence making it in-distinguishable to the model. In this project, the team will focus on developing the model to detect AI-generated text in the semantic level only.

Due to limited computational resources and bandwidth latency, our approach will prioritize the adoption of lightweight architectures. For real-time deployment, the current model is designed to handle small-scale request sizes to ensure efficient processing. While hybrid deployment could accommodate larger request sizes, this aspect will not be addressed in the scope of this project.

Lastly, to avoid submissions to multiple systems for academic integrity checking, the detection model should integrate to the existing platforms. However, as the backend of these systems are not accessible by the public, in this project we will be solely focusing on demonstrating the capabilities of the AI-generated text detection system.