Mining Textual Patterns: A Data-Driven Approach to Distinguish AI-Authored Content

Anirudh Prashant Kalghatkar
Department of Computer Science
University of Colorado Boulder
Anirudh.Kalghatkar@colorado.edu

Sagar Swami Rao Kulkarni
Department of Computer Science
University of Colorado Boulder
SagarSwamiRao.Kulkarni@colorado.edu

Pavan Sai Appari
Department of Computer Science
University of Colorado Boulder
Pavan.Appari@colorado.edu

Sachin Kashinath Rathod
Department of Computer Science
University of Colorado Boulder
Sachin.Rathod@colorado.edu

Nihal Srinivasu
Department of Computer Science
University of Colorado Boulder
Nihal.Srinivasu@colorado.edu

MOTIVATION

The introduction of sophisticated large language models (LLMs) has significantly transformed the generation and accessibility of written text, providing both opportunities and challenges across various domains. These advanced AI systems are capable of producing text that is often indistinguishable from that written by humans, revolutionizing fields such as content creation, customer service, and language translation. However, this technological advancement poses a significant challenge: differentiating between LLM-generated language and human-authored information. This challenge is particularly crucial in academic environments, where the integrity and originality of student work are of paramount importance. The capacity to discern whether an essay was written by a student or created by an LLM is critical for upholding academic standards and ensuring fair grading practices.

As LLMs become more sophisticated, traditional plagiarism detection methods, which typically rely on identifying similarities between texts, are losing their effectiveness. These classic approaches struggle to detect the nuanced and coherent outputs generated by modern LLMs, necessitating the development of more subtle and robust detection algorithms. This research effort addresses the critical need for reliable AI detection methodologies by leveraging a large dataset comprising both student essays and LLM-generated texts. By analyzing this dataset, the initiative aims to identify distinctive features and patterns that can be used to differentiate between human and AI-generated content.

The goal of this initiative is to enhance the tools available for maintaining academic integrity, providing educators and institutions with reliable methods to detect and address the use of AI in student work. By fostering open research and promoting transparency, this project not only aims to improve the detection of LLM-generated content but also contributes to the broader field of AI detection in real-world contexts. The outcomes of this research will be beneficial for instructors and educational institutions, helping to ensure that assessments reflect genuine student effort and learning. Furthermore, this work underscores the importance of adapting to technological advancements while preserving the core values of education and academic integrity.

LITERATURE SURVEY

The rapid advancement of natural language generation technologies, exemplified by OpenAI's ChatGPT and Google's LaMDA, has significantly enhanced the ability of AI systems to replicate human writing styles. These improvements in large language models (LLMs) have made it increasingly challenging to distinguish AI-generated text from human-authored content. This development raises concerns about potential misuse in academic dishonesty, misinformation, and phishing attacks. Consequently, there is a growing interest in developing robust methods for detecting AI-generated text to mitigate these risks.

Several approaches have been proposed to address this challenge. Mitchell et al. (2023) [1] introduced DetectGPT, a zero-shot detection method that analyzes the "probability curvature" of language model outputs to identify potential differences between human and AI-generated text. This approach is promising because it leverages the inherent properties of language models, offering an efficient way to detect AI-generated content without requiring extensive training data specific to each model.

In another effort, Rivera Soto et al.[2] proposed a method that uses writing style representations learned from human-authored texts to detect AI-generated content. This method has proven robust against new language models and can identify the specific language model used for text generation with minimal examples. By focusing on generalizable stylistic markers, this technique demonstrates versatility across different models and datasets. However, challenges remain, such as the long-term effectiveness of these methods against rapidly advancing AI capabilities and the computational costs involved. The need to continuously adapt to advancements in language models and manage the computational load underscores the ongoing research required in this field.

Additionally, frameworks like those discussed by Brundage et al. (2018)[3] highlight the dual nature of AI applications and advocate for collaborative efforts to mitigate risks associated with malicious AI use across various domains. They emphasize the potential security threats posed by AI in digital, physical, and political contexts and call for cooperation among researchers, governments, and the private sector to address these issues. While high-level strategies provide a broad direction, the rapid technological advancements necessitate continuous innovation and adaptation in detection techniques.

This evolving landscape demands more efficient and scalable methods to differentiate AI-generated texts from human-authored ones, ensuring the integrity and reliability of written content across different applications. The development of such methods is crucial not only for maintaining academic honesty but also for protecting the public from misinformation and malicious content. As AI capabilities continue to grow, so too must our efforts to develop sophisticated detection mechanisms that can keep pace with these advancements.

PROPOSED WORK

In our project, we aim to explore various data mining techniques to develop an effective system for distinguishing between AI-generated and human-written text. Our proposed approach includes several key components:

Data Exploration and Preparation: We will begin by conducting extensive Exploratory Data Analysis (EDA) on the "Augmented Data for LLM - Detect AI Generated Text" dataset from Kaggle. This will involve a thorough examination of the dataset to understand its structure, identify any anomalies, and detect potential class imbalances. Addressing class imbalances will be crucial for ensuring that our models are trained on balanced data, which is essential for robust model performance. Additionally, we plan to experiment with different data splitting strategies to ensure that our model development and evaluation are as rigorous and unbiased as possible.

Text Preprocessing: We propose to explore various text preprocessing techniques to prepare the data for model training. This may include traditional methods such as Term Frequency-Inverse Document Frequency (TF-IDF) and Count Vectorization, as well as more advanced techniques like word embeddings. By comparing these preprocessing methods, we aim to identify the techniques that best capture the nuances differentiating AI-generated text from human-written content.

Classification Algorithms: We plan to investigate a range of machine learning algorithms that have shown promise in text classification tasks. This includes traditional algorithms such as Support Vector Machines (SVM) and Logistic Regression, ensemble methods like Random Forest and XGBoost, and deep learning approaches potentially incorporating various neural network architectures. Our goal is to assess the strengths and weaknesses of each algorithm within the context of distinguishing between AI-generated and human-authored texts.

Model Development and Optimization: We will develop a systematic approach for model selection and hyperparameter tuning to ensure optimal performance. This will involve exploring various optimization techniques to enhance the accuracy and efficiency of our models. By systematically evaluating and refining our models, we aim to develop a highly effective system for detecting AI-generated text, contributing to the broader field of AI detection and supporting the integrity of written content in academic and other contexts.

EVALUATION

Our suggested approaches will be quantitatively evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score to ensure a comprehensive assessment of our models.

Accuracy will measure the overall correctness of the model by calculating the ratio of correctly predicted instances to the total number of instances. This metric provides a broad view of the model's performance but may not fully capture its effectiveness, particularly in cases where class imbalances exist.

Precision will evaluate the quality of the model's positive predictions by calculating the ratio of true positive predictions to the total number of positive predictions (both true and false positives). High precision indicates that the model has a low false positive rate, meaning it is adept at predicting instances of a class without mistakenly labeling irrelevant instances as part of that class.

Recall (or sensitivity) will measure the model's ability to identify all relevant instances within a specific class by calculating the ratio of true positive predictions to the total number of actual positive instances (both true positives and false negatives). High recall indicates that the model successfully captures most of the relevant instances, though it may come at the expense of increasing false positives.

F1-Score provides a harmonic mean of precision and recall, balancing the two metrics to give a single measure of performance that accounts for both false positives and false negatives. The F1-score is particularly useful when the data has class imbalances, as it provides a more nuanced view of the model's effectiveness by considering both precision and recall simultaneously.

By employing these metrics, we can comprehensively evaluate the performance of our models. Accuracy will give us a general sense of correctness, precision will inform us about the reliability of our positive predictions, recall will show us how well our model captures all relevant instances, and the F1-score will provide a balanced measure of the model's overall performance in both precision and recall scenarios. This multifaceted evaluation will ensure that our model not only performs well overall but also excels in distinguishing AI-generated text from human-written content across various contexts.

MILESTONES

In **Week 1**, we will conduct a kickoff meeting to assign roles, followed by an in-depth literature review. We will define the project's scope and create a timeline. Additionally, we will collect datasets and set up the necessary tools.

In **Week 2**, we will define and implement feature extraction methods, select initial models, and train and evaluate these models while documenting baseline performance.

In **Week 3,** we will focus on training and comparing advanced models, evaluating and validating them, and conducting error analysis to refine the models.

In **Week 4**, we will conduct the final model evaluation and compare results, create visualizations, compile findings into a final report, and prepare and rehearse our presentation.

Throughout the project, we will have regular check-ins and maintain continuous documentation to ensure smooth progress and thorough record-keeping.

REFERENCES

[1] Mitchell, Eric, et al. "Detectgpt: Zero-shot machine-generated text detection using probability curvature" International Conference on Machine Learning. PMLR, 2023.

[2] Soto, Rafael Rivera, Kailin Koch, Aleem Khan, Barry Chen, Marcus Bishop, and Nicholas Andrews. "Few-Shot Detection of Machine-Generated Text using Style Representations." arXiv preprint arXiv:2401.06712 (2024).

[3] Brundage, Miles, et al. "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation." arXiv preprint arXiv:1802.07228 (2018).