AUTHENTITEXT

Team Members: Lahari Boni, Jayatha Chandra, Sai Akhil Rayapudi

1. Problem Statement:

The increasing prevalence and evolution of AI-generated text underscore the pressing need for reliable tools capable of distinguishing between human and machine-generated content. This distinction is important in preserving the integrity of information circulating in various sectors, including journalism, academia, and online content creation. By leveraging data science techniques, specifically through developing models trained on diverse datasets, we can effectively address this challenge. This project proposes a multi-tiered approach: starting with a single model to establish a baseline, advancing to an ensemble of models, or training a transformer model for improved accuracy and robustness, and finally exploring innovative feature generation techniques to enrich semantic analysis. This multifaceted strategy not only enhances the model's capability to detect subtle differences between human and AI writings but also contributes to the evolving field of natural language processing. Currently, there are similar initiatives in this domain, yet most focus either on simple pattern recognition or on specific textual characteristics, lacking a comprehensive, multi-layered analysis. Our project aims to fill this gap by integrating various methodologies to create a more nuanced and sophisticated detection tool, setting a new benchmark in the field.

2. Dataset Description:

Our dataset comprises about 20,000 essays, some written by students and some generated by a variety of large language models (LLMs). The goal of the competition is to determine whether or not an essay was generated by an LLM.

All the essays were written in response to one of seven essay prompts. In each prompt, the students were instructed to read one or more source texts and then write a response. This same information may or may not have been provided as input to an LLM when generating an essay.

3. Data Modeling and Methodology:

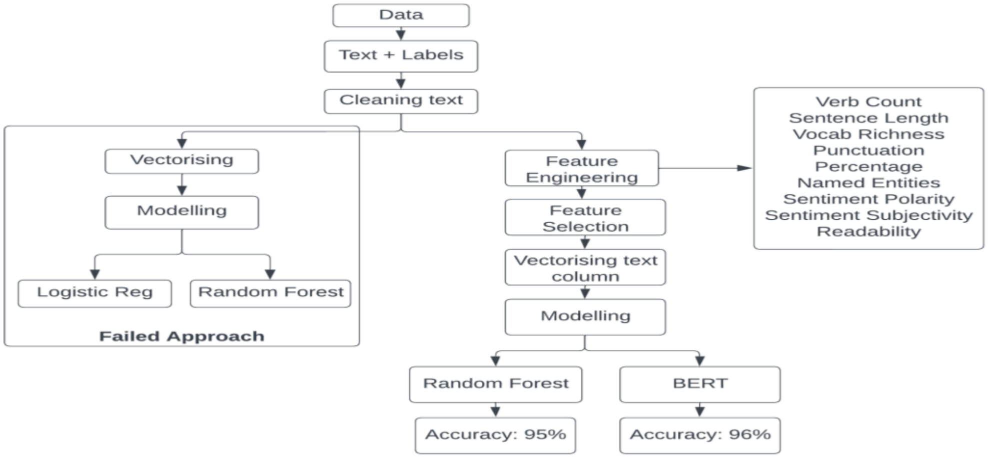


Fig 1: Project Workflow

3.1 Low Risk

In our low-risk approach, we employed logistic regression as one of the classification models to predict if a text is AI-generated or student-written. Logistic regression is a popular method for binary classification tasks, because of its simplicity and interpretability.

During the model training phase, we followed several steps to fit the logistic regression model to our data. Firstly, we preprocessed the data, which included removing stop words, and TF-IDF Vectorization. We then split our dataset into training and testing sets and trained the model to evaluate the model's performance on unseen data.

However, despite our efforts, the logistic regression model failed to achieve satisfactory performance. Upon further analysis, we discovered that there was no linear relationship between the input features and the target variable. Logistic regression assumes a linear relationship between the input variables and the log odds of the target variable, which may not always hold true in real-world datasets.

As a result, the logistic regression model struggled to capture the underlying patterns in the data, leading to poor predictive performance. This highlights the importance of understanding the nature of the data and selecting appropriate models that can effectively capture its inherent complexities.

3.2 Medium Risk

The medium-risk approach in our project seeks to elevate the accuracy and reliability of text classification by employing an ensemble of machine learning models, such as Random Forest and BERT transformer-based models.

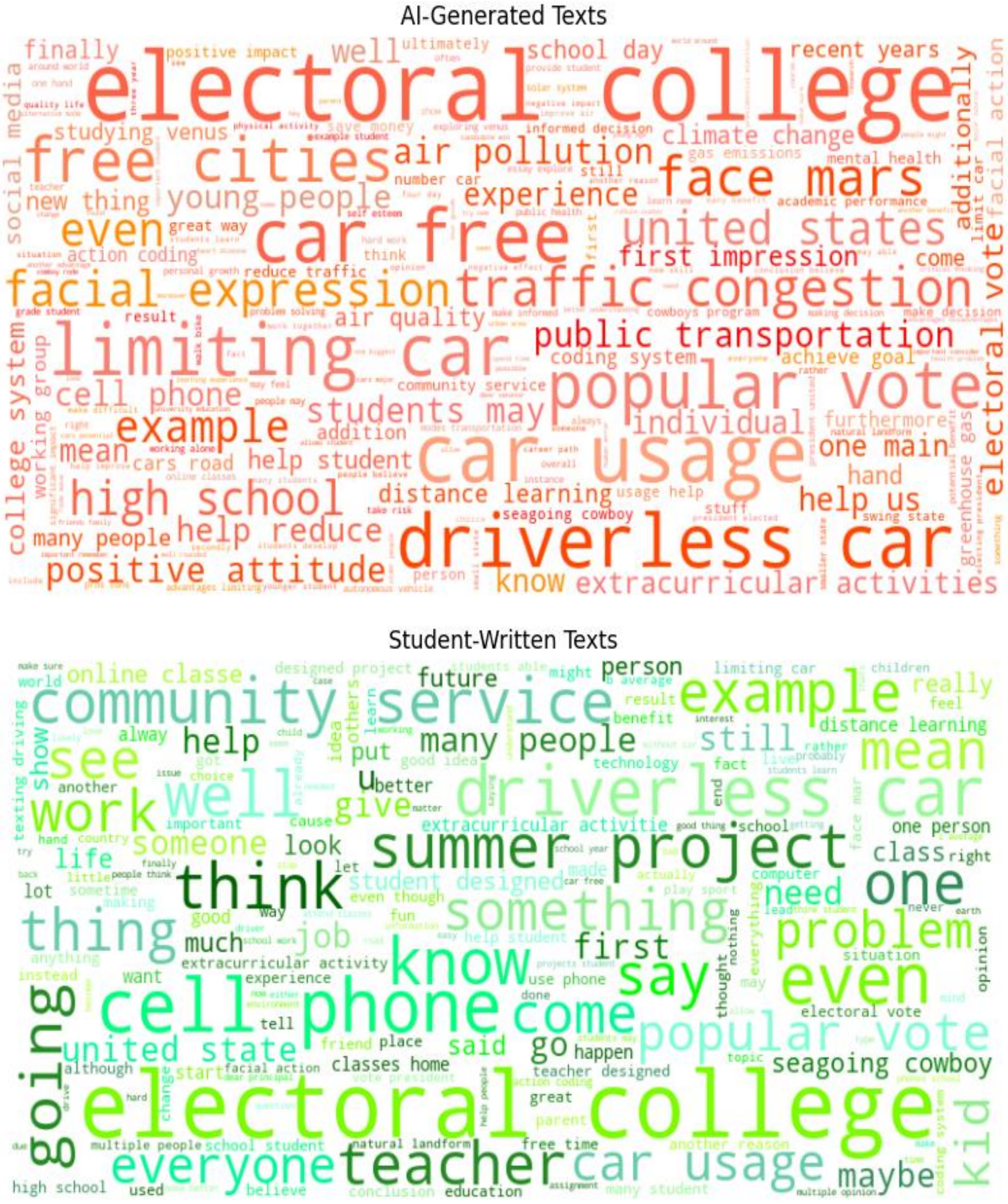
Random Forest: The choice of Random Forest is motivated by their proven efficiency in handling diverse datasets and their ability to reduce overfitting. Here the was just given two features label and text. The model accuracy was 99% which indicated the model was simply memorizing keywords and phrases in the text rather than learning more complex patterns to distinguish human vs AI writing styles.

BERT: The BERT model is selected for its advanced capabilities in understanding context and nuances in natural language processing. The model accuracy was 95% but all the predictions by BERT were biased towards AI text. This indicated the model did not learn properly. Both the models could not completely capture the context of the text, this led to our further approach using feature engineering.

3.3 High Risk

After exploring a simple model using just text and labels, which did not achieve the desired results, we enhanced our approach through comprehensive feature engineering on the text column. Through ANOVA testing for feature selection, we identified critical features that include 'verb_count', 'vocab_richness', 'GPE', 'punctuation_percentage', 'ORG', 'CARDINAL', 'DATE', 'sentiment_polarity', 'readability', 'TIME', 'sentence_length', 'sentiment_subjectivity', and 'cleaned_text’. These features significantly contributed to our model’s performance. By integrating these features into a Random Forest classifier, we substantially enhanced the accuracy of our AI text detection model, achieving approximately 97% accuracy. This method highlights the importance of meticulous feature engineering and selection in crafting more precise and effective models. In addition to refining our model, we also developed a user-friendly interface (UI) to streamline the process of text analysis and prediction. We also included a functionality that dynamically highlights key terms within the essay text, indicating their importance in the model's decision. Leveraging TF-IDF scores, significant words are visually distinguished based on their predicted impact enhancing the interpretability. Furtherly, we integrated plots providing insights into vocabulary richness, verb usage, and punctuation percentage, offering users a comprehensive understanding of the text's linguistic characteristics.

4. Results:



AI-Generated Texts

Student-Written Texts

Fig 2: Word clouds for AI and Student Texts

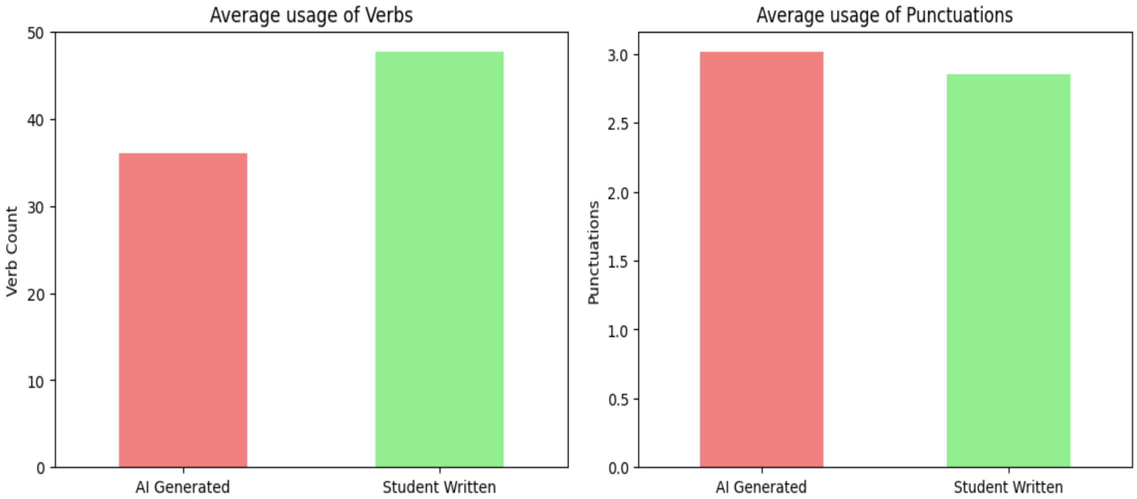
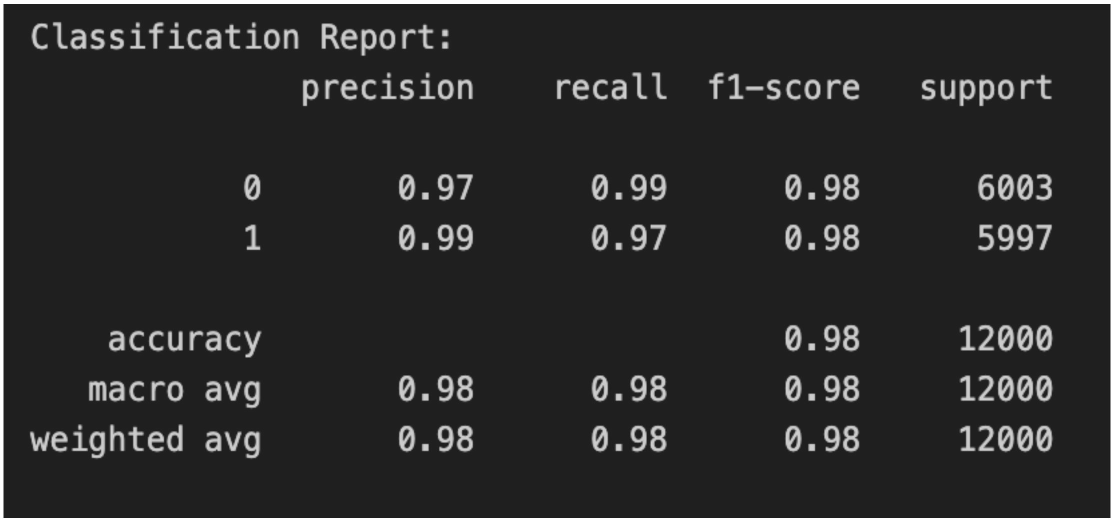


Fig 3: Bar charts for verbs and punctuations usage in AI and Student texts



```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.99      0.98      6003
           1       0.99      0.97      0.98      5997

    accuracy                           0.98     12000
   macro avg       0.98      0.98      0.98     12000
weighted avg       0.98      0.98      0.98     12000
```

Fig 4: Classification report for random forest model

User Interface: AuthentiText app deployment
GitHub: Code

5. References:

1. **https://www.kaggle.com/competitions/llm-detect-ai-generated-text/data**
2. Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). GLTR: Statistical Detection and Visualization of Generated Text. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 111-116). Florence, Italy: Association for Computational Linguistics.
3. Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2021). DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. arXiv preprint arXiv:2105.14123.