GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

PRESIDENCY UNIVERSITY

Private University Estd. in Karnataka State by Act No. 41 of 2013

BANGALORE

PRESIDENCY GROUP
OVER 40 YEARS OF ACADEMIC WISDOM

A Project Report

On

“ AI VS HUMAN – Academic Essay Authenticity Challenge”

Batch Details

Sl. No.	Roll Number	Student Name
1	20211CST0043	V.Sri Manjunath Reddy
2	20211CST0050	Rakshitha M
3	20211CST0002	JDVS Mallika
4	20211CST0029	P Hari Narayana

School of Computer Science,

Presidency University, Bengaluru.

Under the guidance of,

Ms.Sandhya L

School of Computer Science,

Presidency University, Bengaluru

1.Introduction about Project

- This project focuses on addressing the increasing use of AI tools in generating essays, particularly in educational contexts, where maintaining academic integrity is crucial. The aim of the system is to classify essays as either machine-generated or human-authored. The system is capable of identifying texts generated by various large language models, such as ChatGPT, and determining whether the content is AI-generated (Fake) or written by a human (Real). This binary classification task extends to essays written by both native and non-native speakers in two languages: English.

2.Literature Review

Paper Name	Author(s)	Limitations	Advantages	Disadvantages
(The Rise of AI in Content Generation 2019)	Radford	Focused mainly on transformer architecture without addressing practical applications.	Introduced transformer-based architectures that improved text generation.	Limited focus on challenges in academic settings and authenticity.
Challenges in Distinguishing AI-Generated Text (2019)	Zellers	GPT-2 detection model was not robust across diverse texts and contexts.	Highlighted the difficulty in distinguishing between AI-generated and human text.	Struggled with high false positive rates in detection.
Challenges in Distinguishing AI-Generated Text . (2020)	Brown	Detection remains imperfect, especially for advanced models like GPT-3.	Demonstrated significant improvement in text fluency with GPT-3.	The increased fluency of models like GPT-3 complicates detection.
Existing Detection Techniques (2020)	Bakhtin	Detection models like RoBERTa are still limited to certain text structures.	RoBERTa-based detectors show reasonable accuracy in distinguishing AI text.	May not detect subtle, sophisticated AI-generated content.
Existing Detection Techniques . (2020)	Ippolito	The study's models might not generalize well to highly diverse writing styles.	Focused on context, sentence structure, and stylistic patterns for better accuracy.	Models might fail in detecting new or evolving AI writing styles.
Multilingual Detection . (2021)	Elmadany .	AI detection is inconsistent in low-resource languages like Arabic.	Extended AI detection research into multilingual contexts, particularly Arabic.	Limited to specific languages, making global application difficult.
Impact on Academic Integrity . (2021)	Kovačević	Some detection systems still generate false positives, affecting fairness.	Highlighted the growing concern of AI in academic environments.	May result in unfair detection of legitimate student work.

3.Objectives

To address the limitations of existing AI-generated text detection systems, this project proposes a robust method designed to improve accuracy, adaptability, and scalability. The key components of the proposed method are as follows:

1. Use of the Roberta-Base-OpenAI-Detector Model
2. User-Friendly Interface with Dash
3. Adaptive Model Training
4. Login and signup pages with correct connectivity to database

4.Methodology

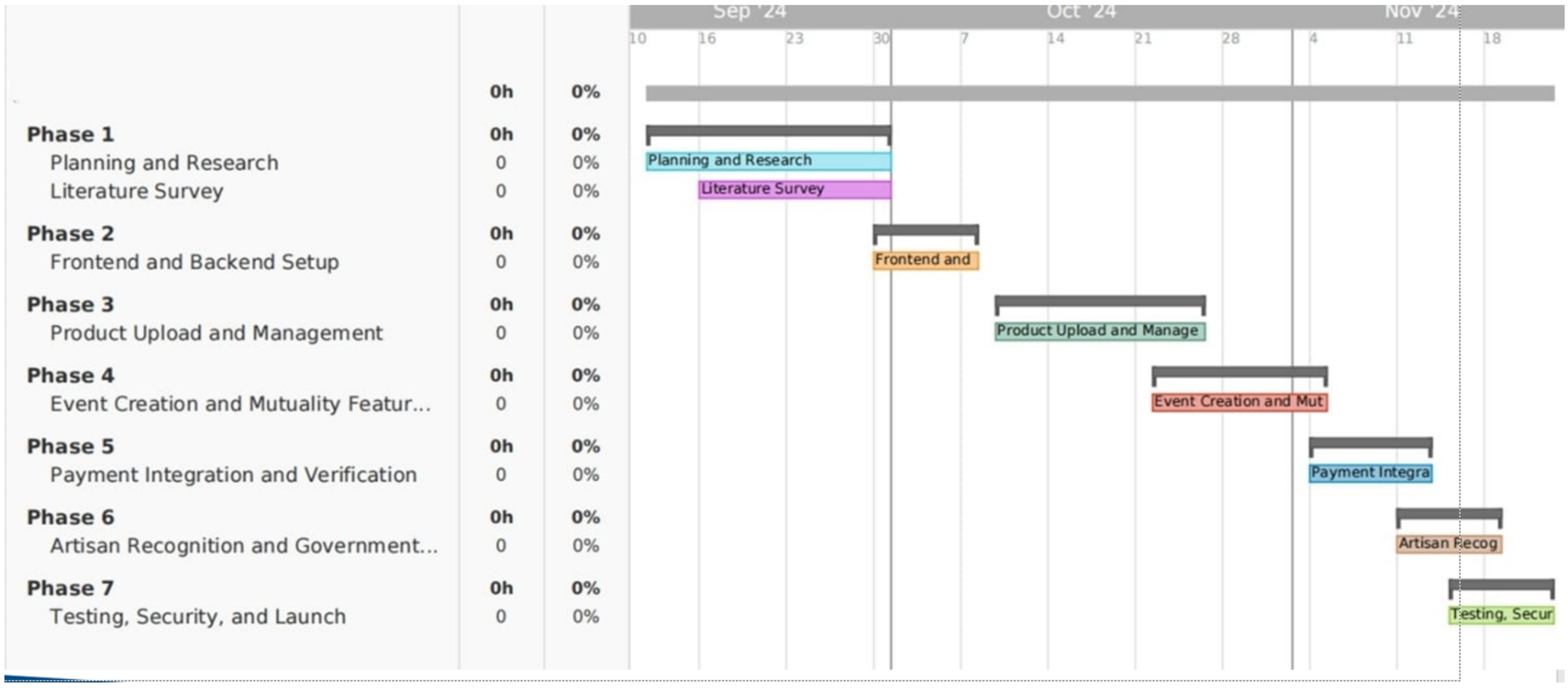
To address the challenge of distinguishing human-authored essays from AI-generated content, a binary classification system was implemented utilizing natural language processing (NLP) and machine learning techniques. The methodology consists of the following steps:

1. **Data Collection** The dataset comprises two distinct sources: Human-authored essays: Essays were sourced from the ETS Corpus of Non-Native Written English. This corpus contains a wide range of writing samples from non-native English speakers, allowing the model to analyze human-authored essays with varying syntactic styles and proficiency levels. The corpus is especially valuable because of its diversity in writing patterns, making it a robust resource for training models. AI-generated essays: Texts generated by seven large language models (LLMs) were included: GPT-3.5-Turbo, GPT-4o, Gemini-1.5, Llama-3.1 (8B), Phi-3.5-mini, and Claude-3.5. Each model provided multiple samples of essays, covering various academic writing prompts. These AI-generated essays simulate responses that might be expected in a university setting.

2. **Preprocessing** Both human and AI-generated essays underwent preprocessing to ensure consistency and remove noise from the text. The following steps were employed: Tokenization: Each essay was tokenized into individual words or subword units to allow for better linguistic analysis. Lowercasing: All text was converted to lowercase to avoid discrepancies due to case sensitivity. Stopword Removal: Common English stopwords were removed to focus on more meaningful content. Lemmatization: Words were lemmatized to their base forms, ensuring that variations of a word (e.g., "writing" vs. "write") were treated consistently.

3. Feature Extraction Several linguistic features were extracted from both human and machine-generated essays, capturing stylistic and structural elements: Term Frequency-Inverse Document Frequency (TF-IDF): This method quantified the importance of individual words in each essay relative to the entire corpus. The TF-IDF values were used to differentiate between common and unique word usage in human and AI essays.. 7 Lexical Richness and Diversity: Measures such as word length, sentence complexity, and lexical variety were used to capture the richness of human writing versus the more repetitive nature of AI text. Embeddings: Word embeddings like GloVe or BERT were employed to capture contextual and semantic nuances in the essays.

5.Timeline for Execution of Project



6.Expected Outcomes

The proposed system is expected to yield several important outcomes that will contribute to safeguarding academic integrity and ensuring the authenticity of written content in educational settings:

1. Improved Detection Accuracy
2. User-Friendly Interface
3. Scalable Solution for Large Institutions
4. Explainable and Transparent AI

7.Conclusion

This project offers a robust solution for detecting AI-generated essays, ensuring academic integrity by distinguishing between human and machine-authored texts. Utilizing the Roberta-base-openai-detector model, it provides accurate classification with support for both English and Arabic essays. The user-friendly interface and explainability features enhance accessibility and transparency, while the system's adaptability ensures it remains effective against evolving AI models. By promoting ethical AI use in education, this tool will help institutions maintain fair academic practices and uphold authenticity in student work.

8.References

[1] R. Corizzo and S. Leal-Arenas, "A Deep Fusion Model for Human vs. Machine-Generated Essay Classification," IEEE, 2023.

[2] W. H. Pan, M. J. Chok, J. L. S. Wong, Y. X. Shin, Z. Yang, Y. S. Poon, C. Y. Chong, D. Lo, and M. K. Lim, "Assessing AI Detectors in Identifying AIGenerated Code: Implications for Education," IEEE, 2024.

[3] V. Pandita, A. M. Mujawar, T. Norbu, V. Verma, and P. Patil, "Text Origin Detection: Unmasking the Source – AI vs Human," IEEE , 2024.

[4] D. Yan, M. Fauss, J. Hao, and W. Cui, "Detection of AI-generated Essays in Writing Assessments," Psychological Test and Assessment Modeling, vol. 65, 2023.

[5] X. Peng, Y. Zhou, B. He, L. Sun, and Y. Sun, "Hiding the Ghostwriters: An Adversarial Evaluation of AI-Generated Student Essay Detection," arXiv:2402.00412v1 [cs.CL], Feb. 2024