AI vs. Human: Academic Essay Authenticity Challenge

**A PROJECT REPORT**

***Submitted by,***

**Lavanya J- 20211CSE0380 DrushtiTR-20211CSE0387 Swarna BB- 20211CSE0371**

**Sonali Siddaling Bangi - 20211CSE0403 Chaya G - 20211CSE0405**

***Under the guidance of,***

**Mr. Jinesh V N**

***In partial fulfillment for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE AND ENGINEERING**

**At**

****

**PRESIDENCY UNIVERSITY**

**BENGALURU**

**DECEMBER 2024**

**PRESIDENCYUNIVERSITY**

**SCHOOL OF COMPUTER SCIENCE ENGINEERING**

**CERTIFICATE**

This is to certify that the Project report **“AI vs. Human: Academic Essay Authenticity Challenge”** being submitted by **Lavanya J(20211CSE0380), Dushti T R(20211CSE0387), Swarna B B (20211CSE0371) , Chaya G (20211CSE0405), Sonali Siddaling Bangi (20211CSE0403)** in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering is a bonafide work carried out under my supervision.

|  |  |
| --- | --- |
| **Mr. Jinesh V N**  **Professionalization** School of CSE Presidency University | **Dr.Asif Mohammed**  HoD  School of CSE & IS Presidency University |

|  |  |  |
| --- | --- | --- |
| **Dr.L SHAKKEERA**  Associate Dean School of CSE  Presidency University | **Dr.MYDHILI NAIR**  Associate Dean School of CSE  Presidency University | **Dr.SAMEERUDDIN KHAN**  Pro-Vc School of Engineering Dean -School of CSE&IS  Presidency University |

**PRESIDENCYUNIVERSITY**

**SCHOOL OF COMPUTER SCIENCE ENGINEERING**

**DECLARATION**

We hereby declare that the work, which is being presented in the project report entitled **AI vs Human: Academic Essay Authenticity Challenge** in partial fulfillment for the award of Degree of **Bachelor of Technology** in **Computer Science and Engineering**,is are cord of our own investigations carried under the guidance of **Mr. Jinesh V N, Assistant Professor, School of Computer Science Engineering & Information Science, Presidency University, Bengaluru.** We have not submitted the matter presented in this report any where for the award of any other Degree.

|  |  |
| --- | --- |
| **Name** | **Roll No** |
| Lavanya J | 20211CSE0380 |
| Drushti TR | 20211CSE0387 |
| Swarna BB | 20211CSE0371 |
| Sonali Siddaling Bangi | 20211CSE0403 |
| Chaya G | 20211CSE0405 |

**ABSTRACT**

The rapid advancements in artificial intelligence(AI)have introduced transformative changes in education, particularly in academic writing. Generative AI tools like ChatGPT are now capable of producing coherent, well-structured, and contextually relevant essays, blurring the lines between human-authored and machine-generated content. This has sparked significant concerns about the authenticity of academic work and the potential erosion of academic integrity. This study, I vs. Human: Academic Essay Authenticity Challenge ,delves into the interplay between human creativity and AI-generated content, examining how these tools are reshaping traditional academic practices. By comparing essays generated by AI with those written by students, the research explores the distinct stylistic, structural, and analytical differences, alongside the challenges in distinguishing between the two.

Theprojectemploysamultidisciplinaryapproach,integratinglinguisticanalysis,AI-detection tools, and subjective evaluations by educators to assess the effectiveness of existing detection methods and the nuances of human versus AI writing.It also examines the ethical implications of AI use in academia, addressing questions of authorship, originality, and the reliance on technology as a shortcut in learning. Preliminary findings reveal that whileAI- generated essays demonstrate high levels of grammatical precision and coherence, they often lack the depth, creativity, and critical reasoning inherent to human writing. Moreover, existing AI-detection technologies show varying levels of success, with many struggling to identify machine- generated content, particularly when it has been edited by humans.

The study highlights the need for institutions to adapt to these challenges by adopting a balanced approach. This includes the development of advanced AI-detection systems, policy reforms to regulate AI usage, and pedagogical strategies to foster critical thinking and originality in students. By addressing these issues pro actively, the research ai ensure that AI serves as a tool for enhancing learning rather than undermining the educational process. Ultimately, this paper contributes to the broader conversation on preserving academic integrity in the age of AI, emphasizing the importance of safeguarding the core

Values of education while embracing the potential of technological innovation.

The rise of AI in academic writing presents not only technological challenges but also profound ethical, psychological ,and institutional implications. This study explore show the accessibility and sophistication of generative AI tools have redefined students' perceptions of effort, creativity, and intellectual growth. Many students view AI as a legitimate resource for enhancing writing skills and addressing language barriers, yet this reliance also risks fostering a culture of dependency that undermines critical thinking and authentic learning experiences. Educators, meanwhile, face the dual burden of identifying AI-generated content and adapting teaching methods to account for the presence of such tools. The project also considers how these developments necessitate a reevaluation of traditional assessment methods, advocating for assignments that demand personalized, experiential, and reflective input—elements less easily replicated by AI systems.

Additionally, the study investigates the evolving role of AI-detection technologies, which, while promising, remain in a constant race against ever-improving AI models. Tools such as stylometric analysis and natural language processing have shown potential but require significant refinement to keep up with the creativity and adaptability of AI-generated content. Beyond technology, the paper underscores the importance of institutional policy changes, emphasizing the need for clear guidelines that define the acceptable use of AI in academic contexts and establish robust measures to uphold academic integrity.

The implications of AI's role in academic writing extend beyond the classroom, touching on broader societal, professional, and technological concerns. This study highlights how the increasing reliance on AI tools could influence the development of essential life skills such as problem-solving, critical analysis, and creative expression. As AI-generated content becomes more prevalent, there is a risk that students might deprioritize these skills, which are foundational to personal and professional growth. Moreover, the ease with which AI can produce polished essays raises concerns about fairness and equity, as not all students have equal access to these advanced tools, creating disparities in academic performance and opportunities.

**ACKNOWLEDGEMENT**

First of all, we indebted to the **GOD ALMIGHTY** for giving me an opportunity to excel in our efforts to complete this project on time.

We express our sincere thanks to our respected dean **Dr. Md. Sameeruddin Khan**, Pro-VC, School of Engineering and Dean, School of Computer Science Engineering& Information Science, Presidency University for getting us permission to undergo the project.

We express our heartfelt gratitude to our beloved Associate Deans **Dr. Shakkeera L and Dr. Mydhili Nair,**School of Computer Science Engineering& Information Science, Presidency University, and Dr. “Dr.Asif Mohammed",Head of the Department, School of Computer Science Engineering& Information Science, Presidency University, for rendering timely help in completing this project successfully.

We are greatly indebted to our guide **Mr. Jinesh V N, Assistant Professor** and Reviewer **Ms. Nithya B A**, School of Computer Science Engineering& Information Science, Presidency University for his/her inspirational guidance, and valuable suggestions and for providing us a chance to express our technical capabilities in every respect for the completion of the project work.

We would like to convey our gratitude and heartfelt thanks to the PIP2001 Capstone Project Coordinators **Dr. Sampath A K, Dr. Abdul Khadar A and Mr. Md Zia Ur Rahman,** department Project Coordinators and Git hub coordinator **Mr. Muthuraj.**

We thank our family and friends for the strong support and inspiration they have provided us in bringing out this project.

|  |  |
| --- | --- |
| **Name** | **Roll No** |
| Lavanya J | 20211CSE0380 |
| Drushti TR | 20211CSE0387 |
| Swarna BB | 20211CSE0371 |
| Sonali Siddaling Bangi | 20211CSE0403 |
| Chaya G | 20211CSE0405 |

**LIST OF FIGURES**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl. No** | **Figure Name** | **Caption** | **Page No.** |
| **1** | **The Methodology Workflow of ai vs human model** | **Demonstrates the methodology pipeline of the too** | **39** |
| **2** | **Output of the project** | **Output screenshot** | **42** |
| **3** | **Timeline of project** | **Timeline** | **43** |

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTERNO.** | **TITLE** | **PAGE NO.** |
|  | **ABSTRACT** | **4** |
|  | **ACKNOWLEDGEMENT** | **7** |
|  | **LIST OFTABLES** | **8** |
| **1** | **INTRODUCTION** | **9** |
| **2** | **LITERATURE REVIEW** | **14** |
| **3** | **RESEARCHGAPSOFEXISTINGMETHODS** | **19** |
| **4** | **PROPOSEDMETHODOLOGY** | **26** |
| **5** | **OBJECTIVES** | **29** |
| **6** | **SYSTEMDESIGNAND IMPLEMENTTION** | **33** |
| **7** | **TIMELINEFOR EXECUTION** | **36** |
| **8** | **OUTCOMES** | **37** |
| **9** | **RESULTSAND DISCUSSIONS** | **41** |
| **10** | **CONCLUSIONS** | **45** |
|  | **REFERENCES** | **46** |
|  | **APPENDICES** |  |
|  | **APPENDIXA:PSEUDOCODE** | **47** |
|  | **APPENDIXB:SCREENSHOTS** | **50** |
|  | **APPENDIXC: ENCLOSURES** | **54** |

**CHAPTER-1 INTRODUCTION**

The rapid advancements in artificial intelligence (AI), particularly in the realm of natural language processing (NLP)and deep learning , have significantly transformed the landscape of text generation. These technologies have enabled the creation of sophisticated AI models capable of producing human-like text, which has led to a myriad of applications, including automated content creation, chatbots, and virtual assistants. However, alongside these innovations, there are growing concerns regarding the potential misuse of large language models (LLMs). This misuse manifests in various forms, such as the dissemination of misinformation, the creation of fake news, and instances of plagiarism in academic contexts.

As AI-generated text becomes increasingly prevalent, distinguishing between human-written andAI-generatedcontenthasbecomeapressingchallenge.TheabilitytoaccuratelydetectAI- generated text is crucial for maintaining the integrity of information and safeguarding against the risks associated with AI misuse. In academic settings, where originality and authenticity are paramount , the stakes are particularly high. The proliferation of AI-generated essays poses a threat to academic honesty, as students may leverage these technologies to produce work that is not their own, undermining the educational process. The urgency of this issue has led to a growing demand for robust algorithms capable of effectively distinguishing between human and AI-generated content. Current detection methods often struggle to keep pace with the rapid evolution of AI text generation technologies, making it increasingly difficult to identify AI- generated text accurately. This challenge is compounded by the diverse writing styles and formats that exist within academic essays, further complicating the detection process. The advent of artificial intelligence (AI) has revolutionized numerous industries, and education is no exception. With the rise of advanced generative AI tools such as ChatGPT, GPT-4, and similar models, students and professionals alike now have unprecedented access to systems capable of producing coherent, persuasive, and seemingly original written content.

These AI- powered platforms, trained on vast datasets of human

language, have the ability to craft essays , solve complex problems, and mimic human creativity at unimpressive level of sophistication. While this technology offers significant opportunities

for improving productivity and learning, it also poses a profound challenge to academic integrity: how do educators, institutions, and society at large differentiate between work authored by human intellect and that generated by artificial intelligence?

This challenge is particularly acute in academic essay writing, a cornerstone of education designedtoassesscriticalthinking,creativity,andtheabilitytosynthesizeinformation.Essays aremeanttoreflectastudent’sintellectualjourney,theirengagementwithsourcematerial,and theirpersonalinsightsonagiventopic.TheemergenceofAItoolscapableofgeneratingessays that often meet or exceed the expected academic standards raises pressing questions about authorship,originality,andtheverypurposeofeducation.IfstudentscansimplypromptanAI to write a polished, well-structured essay in seconds, the traditional metrics for evaluating academic performance are rendered increasingly obsolete.

The authenticity of academic essays is not just an issue of integrity but also one of learning outcomes. Writing an essay is a cognitive process that involves research, analysis, argumentation, and creativity. When this process is outsourced to AI, students may miss out on the opportunity to develop these crucial skills. This reliance on AI has sparked a growing debate within educational institutions: should AI be banned as a threat to genuine learning, or should it be embraced as a tool that can complement human effort? Furthermore, the widespread availability of AI tools has created an uneven playing field. Students with better access to these technologies may gain an unfair advantage, further complicating the issue of equity in education.

Equally concerning is the difficulty in detecting AI-generated essays. Unlike traditional plagiarism, which involves copying existing text, AI-generated content is novel and does not directly replicate any pre-existing material. This renders traditional plagiarism detection tools ineffective, prompting the development of new AI-detection technologies. However, even the most advanced detection tools face limitations,as they often struggle to keep up with the rapid evolution of AI models. Moreover, when students edit AI-generated content to add their personal touch, detection become seven more challenging, raising the stakes for educators and institutions trying to uphold academic standards. The ethical dimensions of this issue are far- reaching. The use of AI in academic writing raises critical questions about the boundaries of acceptable assistance. At what point does using AI transition from being a legitimate aid to outright academic dishonesty? Should students disclose their use of AI tools, and if so, how

Can institutions enforce transparency Additionally, the normalization of AI usage in academic settings could fundamentally alter perceptions of originality and creativity, potentially devaluing the intellectual efforts of human authors. This study, I vs. Human: Academic Essay Authenticity Challenge, seeks to address these pressing concerns by exploring the complex relationship between human creativity and machine-generated content in academic writing. By comparing essays produced by students with those generated by AI, this research aims to identify the stylistic, structural, and analytical differences between the two. It also examines the effectiveness of current AI-detection tools and their potential to up hold academic integrity. Beyond technical solutions, this study delves into the ethical and psychological implications of AI usage in academia, highlighting the motivations behind students' reliance on these tools and the challenges educators face in fostering originality.

Furthermore, this study takes a proactive approach to addressing the broader implications of AI in education. It explores how institutions can adapt to this new reality by developing innovative pedagogical strategies, such as designing assignments that emphasize personal reflection, creativity, and real-world application elements that AI cannot easily replicate. It also advocates for the integration of AI literacy into curricula, empowering students to use these tools responsibly and effectively while understanding their limitations and ethical considerations.

Ultimately, this research aims to contribute to the global conversation about the role of AI in education, emphasizing the need for balanced approach that leverages the benefits of AI while safeguarding the core principles of academic integrity and learning. The challenge of distinguishing between human-authored and AI-generated essays is not just a technical problem—it is aphilosophical,ethical,and practical issue that demands coordinated response from educators, technologists, policy makers, and students. By addressing this challenge head- on, this study seeks to ensure that AI becomes a tool for enhancing education rather than underminingitsfoundationalpurpose.Theimplicationsofthischallengestretchfarbeyondthe academic environment, influencing how society perceives intellectual effort, creativity, and authenticity in an increasingly AI-driven world. Academic institutions have Historically been gate keepers of knowledge, fostering the critical thinking and problem-solving skills necessary for societal advancement. However, with the pro liferation of AI tool scapable of replicating or even surpassing human output sincer tainare as, the fundamental role of

education must be reconsidered. Are we preparing students to think independently, or are we inadvertently encouraging them to depend on AI systems for intellectual labor This dilemma underscores the need for a re-evaluation of educational priorities and practices in the age of artificial intelligence. One significant concern lies in the potential long-term effects of AI dependency on student development. Writing essays, for example, is not merely an academic exercise but a cognitive process that enhances critical reasoning, creativity, and the ability to synthesize complex ideas. When students rely on AI to bypass these intellectual challenges, they risk undermining their capacity to think deeply and independently. Moreover, the convenience of AI-generated content may discourage students from engaging with primary research, exploring multiple perspectives, and developing original arguments—all of which are essential skills for personal and professional growth.

At the institutional level, the widespread use of AI in academic writing introduces new complexities for educators. Traditional methods of assessment, such as essays and written reports, are increasingly vulnerable to manipulation by AI tools, forcing educators to rethink how they evaluate student performance. Assignments that rely heavily on generic prompts or formula ai responses are especially susceptible, as they can be easily completed by AI systems. This necessitates the creation of more innovative and personalized assessments that emphasize critical thinking, creativity, and real-world application—areas where human intellect has a clear advantage over machines.

The broader societal implications of this challenge are equally significant. If AI-generated contentbecomesthenorminacademicandprofessionalsettings,itcoulderodetrustinwritten work, devaluing the authenticity and intellectual effort traditionally associated with human authorship. Employers, for instance, may begin to question the credibility of academic qualifications, particularly if they cannot distinguish between candidates who have genuinely mastered their fields and those who have relied heavily on AI assistance. This could have a ripple effect on industries that depend on original thinking, ethical judgment, and creativity, leading to a potential devaluation of human ingenuity.

Furthermore, the ethical questions surrounding AI use in academia demand urgent attention. Is it ethical for students to use AI to complete assignments ,even if they edit or refine the output Should institutions require students to disclose the extent of AI involvement in their work These questions highlight the need for clear ethical guidelines and institutional policie.

define acceptable and unacceptable uses of AI. At the same time, the development of such policies must account for the diverse cultural,educational,and technological contexts in which AI is used, ensuring that they are both equitable and adaptable.

Another key aspect of this challenge is the arms race between generative AI and detection technologies. As AI models continue to improve, they are becoming increasingly difficult to detect, even for advanced AI-detection tools. This ongoing competition creates a significant burdenforinstitutions,whichmustcontinuallyinvestinnewtechnologiesandtrainingtokeep pacewithAIadvancements.Moreover,therelianceondetectionsystemsraisesconcernsabout privacy and surveillance, as students ‘work maybe subjected to extensive scrutiny in an effort to ensure its authenticity.In response to these challenges,this study emphasizes the importance of fostering AI literacy among both students and educators. By teaching students how to use AI responsibly and effectively, institutions can help them harness the potential of these tools whilemaintainingacommitmenttooriginalityandethicalbehavior.Similarly,educatorsmust be equipped with the knowledge and resources to integrate AI into their teaching practices, creating a learning environment that balances technological innovation with traditional academic values.

**CHAPTER-2 LITERATURE SURVEY**

This paper provides a detailed survey of the challenges and potential indetecting AI- generated text. The authors emphasize the rapid advancements in large language models (LLMs) like GPT and ChatGPT, which produce human-like text indistinguishable from genuine human authorship in many cases. The survey explores various detection methods, categorizing them into statistical, rule-based, and deep learning techniques. Ghosal et al. highlight how adversarial robustness and scalability remain pressing challenges in detection. Oneof the most significant contributions of the paper is its exploration of ethical dimensions, such as the implications of undetected AI-generated text in misinformation and academic dishonesty. Additionally, the paper discusses benchmarks and evaluation metrics used to measure detection performance, arguing for standardized datasets to improve cross- comparability. The authors conclude by proposing the need for continuous monitoring and adaptive systems to keep pace with generative AI advancements[1].

Sadasivan et al. delve into the reliability of current AI-generated text detection systems. The paperbeginsbyoutliningtherapidevolutionofgenerativemodelsandtheirabilitytoproduce increasingly sophisticated text. A core focus is on the inherent challenges faced by detectors, such as their dependence on training datasets and vulnerability to adversarial inputs.The authors conduct empirical experiments to evaluate the generalizability of detection systems across different domains and languages. They find that many systems, while effective on specificdatasets,failtoperformwellindiversereal-worldscenarios.Thepaperalsohighlights thelimitationsofpurelydata-drivenapproachesandadvocatesforhybridsystemsthatcombine linguistic insights with deep learning. This study underscores the need for detectors to evolve alongside generative models to ensure reliability in dynamic environments[2].

Wu et al. provide a comprehensive survey that highlights the necessity of detecting text generatedbylargelanguagemodels.Thepaperidentifieskeyapplicationareaswheredetection is critical, such as combating fake news, preventing plagiarism, and ensuring authenticity in digital communication.The authors classify existing methods into supervised, unsupervised, and hybrid approaches,evaluating their strengths and weaknesses.Supervised methods excel

Inspecific contexts but require labeled datasets,while unsupervised approaches offer flexibility but often lack precision. Hybrid methods, which combine the strengths of both, are identified as the most promising direction. The paper also discusses future research avenues, including the development of cross-model detectors and the role of explainable AI inenhancing detection transparency.Wuetal.emphasize that tackling adversarial robustness and domain adaptability will be crucial in the coming years[3].

Wang et al. investigate the use of BERT-based deep learning techniques for detecting AI- generated text. The paper details the architecture of BERT and its ability to model complex linguisticpatterns.Theauthorsconductexperimentstofine-tuneBERTforclassificationtasks, achieving high accuracy rates on benchmark datasets.However, the paper also addresses limitations, such as the model's dependency on labeled data and susceptibility to adversarial attacks. To overcome these challenges, the authors propose ensemble methods that combine BERT with other deep learning techniques. They also explore lightweight BERT variants to address computational efficiency, making detection feasible for real-time applications. This workdemonstratesthepotentialofdeeplearningalgorithmsinadvancingdetectiontechnology but also acknowledges the need for continual adaptation as generative models evolve[4].

Bakhtin et al. present an early investigation into detecting machine-generated text, setting the stage for later research in the field. The study introduces a supervised learning-based framework to train models for discrimination tasks. The authors experiment with various datasets and model architectures ,highlighting the challenges of scalability and adaptability as generative technologies evolve.Although the study lacks the sophistication of modern approaches ,itremainsa seminal work that underscores the need for robust detection systems. Theauthorsadvocateforearlyinterventionsinthearmsracebetweengenerativeanddetection technologies, emphasizing the importance of understanding the linguistic and structural patterns unique to AI-generated content[5].

This paper introduces Scarecrow, a frame work for scrutinizing the quality and detectability of AI-generated text. Dou et al. focus on linguistic and contextual inconsistencies, identifying subtle patterns that distinguish machine-generated text from human-authored content.

The frame work evalua test extacross multiple dimensions,such as coherence,fluency,and

factual consistency, providing a holistic approach to detection. The authors conduct extensive experiments with GPT-3 outputs, revealing vulnerabilities in current detection systems. They argue for the adoption of multi-dimensional evaluation metrics to enhance detection accuracy and robustness. Scarecrow is presented as a flexible tool that can be adapted to various detection scenarios, offering insights into the evolving capabilities of generative models[6].

Li et al. propose the MAGE framework, designed to detect machine-generated text in real- worldapplications.Thepaperhighlightstheimportanceofdomainadaptabilityandrobustness in detection methods. By leveraging a combination of supervised and semi-supervised techniques, the authors achieve improved detection performance across diverse datasets, including social media and academic writing. A key contribution of this study is its focus on practical challenges, such as the lack of labeled data in certain domains and the need for lightweight models for deployment in resource-constrained environments. The authors also explore the integration of linguistic features with deep learning methods, offering a hybrid approach that balances accuracy and efficiency[7].

Bahadetal. explore the potential offline-tuningpre-trained language models for distinguishing AI-generated text from human-authored content. The paper evaluates various fine-tuning strategies, comparing their performance across datasets with differing characteristics.The authors demonstrate that fine-tuned models can achieve high detection accuracy but face challenges in terms of generalizability and computational cost. To address these issues, they propose the use of transfer learning and domain-specific adaptations. The study also highlights the importance of evaluating models on diverse datasets to ensure robustness. This work provides valuable insights into the trade-offs between computational efficiency and detection accuracy, advocating for balanced approaches in real-world applications[8].

Petropoulos and Petropoulos evaluate the effectiveness of RoBERTa and Bi-LSTM architecturesindetectingAI-generatedtext.RoBERTa,withitsrobustpre-trainingonadiverse corpus, achieves superior performance, particularly in scenarios requiring high precision. Bi- LSTM models, on the other hand, offer competitive results in resource-constrained environments, making them suitable for light weight applications.The authors propose a hybrid detection framework that integrates RoBERTa and Bi-LSTM, combining their strengths to improve overall performance.They also highlight the importance of explain ability indetection

systems, emphasizing the need for transparent models that can provide insights into their decision-making processes. This study underscores the potential of combining advanced architectures with lightweight solutions to address the diverse requirements of AI-generated text detection [9]

**CHAPTER 3**

**RESEARCH GAPS OF EXISTING METHODS**

# AI Detection Methods

Inconsistent Detection Accuracy: Existing AI detection tools may struggle to accurately distinguish AI-generated content from human-written essays, especially as AI technology evolvesandimproves.Theremaybegapsindetectionaccuracyacrossdifferentwritingstyles or types of AI models.

Contextual Understanding: Current methods might not fully capture the nuanced, context- dependent nature of academic writing. AI-generated content often mimics human-style text, but methods may miss subtler indicators of in authenticity in complex academic arguments.

Model-Specific Limitations: Different AI models (GPT-3, GPT-4, or other LLMs) generate text differently. Current detection tools may be tailored to specific models but struggle when confronted with newer or less studied ones.

# Human Writing Variability

Understanding Human Diversity in Writing: Human-written essays vary widely based on factors like educational background, writing experience, cognitive abilities, and even cultural influences.Existingdetectionmethodsoftenfailtoaccountforthisvariability,leadingtofalse positives or negatives in authenticity checks.

Identification of Genuine Human Errors: AI systems can exhibit unnatural or overly polished writing, but human writers can also produce poor grammar or inconsistencies, making it hard to distinguish between genuine human flaws and AI-generated issues.

**Ethical and Privacy Concerns** PrivacyofStudentData:Currentresearchmethodsmaynotadequatelyaddressconcernsabout privacy when scanning academic essays for authenticity, particularly in using AI-based tools that analyze and store student data.

FairnessandBias:Detectionmethodscouldinadvertentlyintroducebias,forexample,favoring one type of writing over another based on cultural or linguistic factors.

There’s a gap in ensuring fairness in AI vs. human authenticity detection, especially for non- native English speakers or students with disabilities.

# Real-Time Detection and Response

Time Constraints in Detection\*\*: Most current systems focus on post-submission analysis of essays, but there is a gap in real-time detection that could alert educators when students are using AI to write or assist with their essays. This gap is significant for preventing academic dishonesty in an environment where AI writing assistants are becoming more accessible.

# Academic Integrity and AI's Role in Education

AI-AssistedLearningvs.Cheating:There’sinsufficientresearchintohowAItoolsareusedin learning environments where students use AI as an aid rather than as a crutch. Identifying ethical boundaries between AI assistance and academic dishonesty is a key challenge.

Educational Systems’ Adaptation: There’s a need for more research into how educational systems can adapt to the presence of AI tools. This includes revisiting what constitutes academic dishonesty and establishing guidelines for AI usage in academic settings.

# Improving AI Writing Authenticity

AI Text Style Mimicry: While AI can generate human-like text, there’s still a gap in methods that can ensure AI writing isn’t overly formulaic or lacks true creative originality. Research could focus on enhancing AI’s ability to mimic diverse academic writing styles convincingly while retaining the essential features of originality.

Cross-Linguistic Challenges: Existing methods predominantly focus on English-language essays, and there’s a research gap in detecting AI-generated content in non-English academic writing, especially as AI tools become more sophisticated and multilingual.

# Impact of AI Writing on Education

Long-Term Effects of AI Use in Academic Writing:There’s limited research on the long-term impact of AI-assisted writing on students' learning processes and critical thinking skills. Understanding how AI affects writing development,creativity,and academic growth is an area that needs deeper exploration.

Byexploringthesegaps,yourprojectcouldcontributevaluableinsightsintoimprovingthe

Authenticity challenge in academic writing and foster a better understanding of the role plays in education.

# AI DetectionTools for Hybrid Writing

Human-AI Collaboration: Many students use AI tools in conjunction with their own ideas, creating "hybrid" essays that combine human thought and AI-generated content. Current detectionmethodsaregenerallygearedtowardfullyAI-generatedorfullyhuman-writtentexts, butthereisagapinaccuratelyidentifyinghybridessays.Researchintodistinguishingbetween human and AI contributions within the same piece of writing is needed.

Tools for AI-Assisted Editing:There is little research on tools that specifically detect instances where AI has been used for editing or refining human-written drafts, rather than generating content from scratch. Developing algorithms to spot these subtle changes would be important to fully address the authenticity challenge.

# Enhanced AI Literacy

Educating Students and Educators About AI: As AI becomes more integrated into academic life, there's a need for research into how well students and educators understand AI's role in writing. Research could focus on the gaps in AI literacy and the development of resources for ethical AI use in academia.Without a strong understanding ofAI's capabilities and limitations, students and educators may either misuse or fail to properly address AI-generated content. theAge of AI:Educational systems could explore how writing instruction might change in response to AI tools. There’s a research gap in understanding how writing curriculashouldevolvetohelpstudentsbothavoidAI-assistedcheatinganduseAIeffectively for brainstorming, drafting, and refining ideas.

# Improved AI Writing Models

AI's Understanding of Context and Tone: Although AI has made significant strides in generating coherent and contextually appropriate essays, it still struggles with tone, subtlety, and deeper thematic understanding in academic writing.

AI and Citation Practices: AI-generated essays often misuse citations or generate fabricated references. Research into improving AI's ability to accurately reference and cite sources in a

way that adheres to academic integrity standards is another key gap. Custom AI Writing Models:There may be agapin research on custom-tailored AI models that canmimic specific academic disciplines' writing styles. Academic writing in the humanities, social sciences, and natural sciences differs significantly,andAI tools may need to be specialized for these contexts.

# AI-Generated Text Evaluation Metrics

Authenticity and Creativity Measurement: Evaluating the authenticity of writing involves not only detecting whether a text was AI-generated but also assessing the creativity, originality, andcriticalthinkingdemonstrated.Thereisagapincreatingstandardizedmetricsforassessing AI writing that account for originality in content creation, which is central to academic integrity. AIOutputvs.HumanWritingFeedback:Currentresearchmightnotfullyexplorehowhumans engagewithAI-generatedtextsaftertheyhavebeengenerated(e.g.,providingfeedback,edits, or reworking AI suggestions). Exploring how AI writing tools influence human revisions or ideas is an area for further research.

# Cross-Cultural Perspectives on AI Writing

Cultural Bias in AIContent: AImodels are often traine don large datasets that may reflect the cultural, social, and linguistic biases of the authors and sources they learn from.There’sa gap in understanding how AI writing tools might produce content that is culturally biased or misrepresents certain global perspectives. This gap has implications for academic integrity, especially in diverse educational settings where students from different backgrounds may use AI tools.

Cultural Variability in Academic Writing: Different academic traditions emphasize distinct writingstyles,argumentation approaches,and citation formats.AI-generated essays might not always align with these culturally specific conventions,which can complicate the detection of authenticity and raise questions about AI’s role in cross-cultural academic contexts.

# Impacton Assessment and Grading

Adapting Grading Systems to AI Use: One significant challenge in the academic world is adaptingassessmentpracticestoaccountforAItools.Thereisagapinresearchonhowgrading

systems can be modified to reflect students' use of AI—whether it's as a research aid, a brainstorming tool, or a text generator. Can traditional grading methods remain valid, or is there a need for new, more holistic evaluation methods?

AssessingProcessvs.Product:InanAI-assistedworld,it'spossiblethatmoreemphasiscould beplacedonthewritingprocessratherthanthefinalproduct.Thisresearchgapcouldexplore how to evaluate a student's development, thinking, and understanding throughout the writing process (from initial idea to final submission), making the detection of AI-generated content less critical in an assessment framework.

# AI Writing Assistants and Academic Integrity Policies

Universities' Response to AI Tools: While some universities are already establishing clear policies regarding AI in academic work, there remains a gap in understanding how well these policies are communicated and enforced. Research could explore how institutions define the boundaries of AI use in academic writing, and whether students are fully aware of these policies.

Developing New Integrity Guidelines: There’s an opportunity to create new guidelines that define what constitutes academic dishonesty in the context of AI-assisted writing. Research could investigate how these guidelines might differ between disciplines (e.g., humanities vs. technical fields) and how they can be implemented fairly and effectively.

# Long-Term Effects on Academic Authorship

Shifting Perceptions of Authorship: With AI’s growing presence in academic writing, there could be a shift in how authorship is perceived. Is the work of a human writer still valid if AI tools played a significant role? There’s a gap in exploring how this shift may affect academic careers, publication standards, Ana the future of authorship in scholarly fields

**CHAPTER-4 PROPOSED METHODOLOGY**

The proposed methodology for investigating the academic essay authenticity challenge posed byAIadoptsarobust,multi-facetedapproachtoensureathoroughexplorationofthetopic.A mixed- methods design, combining both qualitative and quantitative research, will provide a comprehensiveunderstandingofhowAItoolsarebeingutilizedinacademicwritingandtheir implicationsforauthenticity.Tobegin,surveysandquestionnaireswillbe distributed to three primary groups: students, educators, and academic policymakers. These surveys will capture data on the frequency of AI tool usage, the motivations behind relying on AI for academic work, levels of awareness regarding institutional policies on academic integrity, and perceptions of how AI impacts the quality and authenticity of essays. This data will reveal usage trends, behavioral patterns, and attitudes toward AI in academic settings, providing a foundation for deeper analysis.

To build on these insights, focus group discussions will be conducted with students and educators.ThesediscussionswilldelveintothenuancedethicaldilemmassurroundingAIuse in academia, including concerns about dependency on AI, the potential erosion of critical thinking skills, and the challenges of fostering authenticity in a landscape increasingly influenced by technology. Educators will be encouraged to share their experiences in identifyingAI-generatedessaysandtheirperspectivesonadaptingteachingmethodstoaddress this challenge,while students will discuss their motivations for using AI tools and the pressures they face in academic environments.The focus groups aim to uncover not only the challenges but also potential solutions for integrating AI ethically into academic practices.

In addition to gathering qualitative insights, a comparative essay analysis will be undertaken to evaluate the tangible differences between human-written essays and those generated or assisted by AI tools. A sample of essays will be collected anonymously, categorized by their creation process (human-only vs. AI-assisted), and assessed by academic experts. Essays will be evaluated based on criteria such as originality,depth of analysis,coherence,argumentation, and alignment with academic writing standards. This analysis will help identify specific markers of AI-generated content and provide a clearer understanding of how thes essays differ from authentic human work.

This research’s main goal is to thoroughly assess the Bidirectional Encoder Representations

from Transformers (BERT) model's capacity to distinguish between text authored by people and machines—a distinction that is becoming more and more important in maintaining the accuracy of information on the internet. As part of our technique, we prepare a variety of datasets with varying text difficulties, ranging from straightforward automated responses to sophisticated narrative texts. To meet the requirements of BERT's analysis, we will do preprocessing processes like tokenization and normalization. After that, we will run several experiments to test the model in controlled, blind, and edge case settings. Metrics including accuracy, precision, recall, and F1-score will be used to assess BERT's performance; special attentionwillbepaidtothemodel'sprocessingefficiencyandclassificationaccuracy.Ourgoal is to pinpoint the model's advantages as well as any potential drawbacks, especially when it comes to managing complex texts. This thorough investigation, which provides insights into the applicability and dependability of BERT in AI text detection scenarios, will make a substantial contribution.

# Dataset Preparation

There are 29,145 entries in the training data, which may be accessible in a CSV file.Eachentry is identified as either AI-generated or human-written. The dataset consists of two columns: "generated," a binary indicator where "0" indicates material created by humans, and "text,"

whichcontainstheessay'scontent.Thisdatasetensuresthoroughtrainingandevaluationofthe

BERT model for precise AI text detection by offering a well-balanced mix of themes. The dataset's wide range of topics enables a thorough evaluation of the model's capacity to distinguish between information generated by humans and machines in variety of scenarios.

# Data Preprocessing

One of the most important steps in getting the dataset ready for BERT model training is data pretreatment.Tomakesurethetextdataisclear,consistent,andappropriateformodeltraining, this procedure entails a number of crucial steps. To preserve consistency, text normalization methods like changing every character to lowercase and eliminating special characters and punctuation are first used. Tokenization is a necessary step for BERT, which uses tokenized inputs to work,as it divides the text into individual words or subwords.To cut down on noise, stop words—common words that don't really contribute anything to the text's semantic meaning—are eliminated. To guarantee that every text sequence is the same length—a

requirement for batch processing during model training—padding and truncation are also used.

# Data Visualization

Understanding and interpreting the dataset requires the use of data visualization, which sheds light on its distribution and structure. First, we use heat maps to see if any missing value in the training and testing sets. Any null values that need to be addressed during preprocessing are visible in these heat maps. Next, pie charts are used to show the label distribution in both the training and testing sets, emphasizing the percentage of texts created by AI and texts produced by humans. The dataset's balance is verified by the sevisuals,which is necessary for objective model training.

# Model Training and Validation

A methodical approach was used for the training and assessment of the BERT model, starting with data preprocessing, which included tokenization, padding, and truncation to standardize the input data format appropriate for BERT. The model was then adjusted on this prepared dataset, effectively adjusting the model's parameters with the use of the Adam optimizer and the cross-entropy loss function. The pre-trained BERT was fine-tuned to the goal of telling machine-generated text from hand-written material. A thorough assessment of the model's performance was conducted using precision, recall, F1-score, and accuracy. This allowed for the identification of potential areas for development as well as a complete evaluation of the model's ability to effectively classify the text. This all-encompassing strategy made sure the BERT model was optimized for efficacy and efficiency.

**CHAPTER-5 OBJECTIVES**

# Detection and Prevention

AI-Driven Detection Tools:

Leverage advanced AI-based detection systems(e.g.,GPTZero,TurnitinAIdetection)to identify AI- generated content in academic submissions.

Continuous Improvement of Detection Algorithms:

Advocate for the regular updating of detection systems to keep pace with evolving AI capabilities, ensuring consistent accuracy.

Proactive Screening Processes:

Introduce plagiarism checks that assess not only verbal time copying but also AI-aided paraphrasing and synthesis.

Educational Workshops:

Train educators and students on identifying signs of AI-written work and understanding the risks of using AI unethically.

# Assistance without Overreliance

Guidelines for AI Use:

Develop clear guidelines for acceptable AI usage in academic settings, such as using AI for brain storming, grammar checking, or feedback, rather than full essay generation.

Skill-Building Exercises:

Create assignments that challenge students to combine AI assistance with their own critical thinking, such as comparing AI suggestions with their original ideas.

AI as a Mentor:

Frame AI tools as digital mentors that support refinement and learning, rather than producing end products.

Focus on Process, Not Product:

Require students to document their writing process, including where and how they used, to encourage reflection and active learning.

# Transparency and Accountability

Mandatory Disclosure Statements:

Require students to include a declaration in the ire says detailing the extent of AI involvement, ensuring transparency.

AI Usage Logs:

Encourage the use of AI tools that log user interactions, allowing educators to review how AI was utilized in the essay-writing process.

Consequences for Misuse: DefineclearconsequencesforunethicalAIuse,rangingfromreducedgradestoacademic penalties, depending on the severity.

Promote Academic Honesty:

Launch campaigns to instill the value of originality and ethical conduct among students.

# Preservation of Academic Integrity

Updated Academic Policies:

Revise academic integrity policies to explicitly address the use of AI in assignments and its ethical boundaries.

Case Studies on AI Misuse:

Incorporate real-world examples of unethical AI use to educate students about its consequences. Zero-Tolerance Policies:

Enforce strict penalties forde liberately submitting AI-generated work without proper disclosure. Institutional Accountability:

Empower institutions to monitor AI-related trends and continuously adapt their policies.

# Developing Critical Thinking

Critical Analysis Assignments:

Assign essays that required analysis and synthesis,whereAI-generated content would struggle to meet the required intellectual rigor.

Debate AI Outputs:

Ask students to critically assess the limitations or inaccuracies oaf-generated responses, promoting analytical skills.

Reflection Papers:

Incorporate assignments where students reflect on how AI influenced their learning process and what they gained from their efforts.

Hands-On Activities:

Encourage activities like peer review so in-class discussions to prioritize human in sights over AI-driven conclusions.

# Balancing AI Usage and Originality

Limits on AI Contributions:

Introduce caps on the percentage of AI-generated content allowed in an essay, ensuring the majority of the work reflects students’ own voices.

Weight Process Over Outcome:

Allocate a portion of grades to the writing process, ensuring students engage with their ideas rather than relying solely on polished AI outputs.

Personalized Essay Topics:

Design essay prompts tailored to students ‘unique experiences, making it harder for AI to generate relevant and original responses.

Promote Creativity:

Foster creativity by encouraging students to experiment within wide as and personal interpretations, which AI struggles to replicate authentically.

# Originality: An Essential Element of Academic Integrity

Originality stands as a fundamental characteristic of academic writing, presenting significant challenge for artificial intelligence. Although AI can produce essays based on given prompts and input data, it predominantly relies on existing knowledge,paraphrasing,and recognizing

patterns.Consequently,contentgeneratedbyAIismorepronetobeingderivativeor,attimes, unintentionally plagiarized, as it may repeatedly draw from the same sources or concepts. In contrast, human authors typically utilize personal insights, creative interpretations, and direct engagement with research, resulting in a distinctive contribution to the academic discourse.

The risk associated with AI-generated essays lies in the potential for students or writers to exploit these tools to circumvent the necessity of original thought. While AI demonstrates proficiency in generating syntactically correct content, it does not inherently produce innovativeideasorofferfreshperspectivesonasubject.Theprocessofessaywritinginvolves not merely restating information but also critically engaging with it—an aspect that AI finds challenging to authentically replicate.

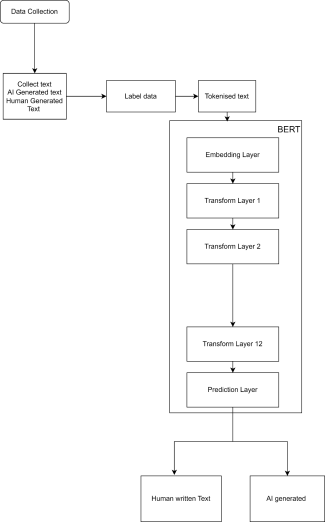
# Intellectual Engagement: A Distinctive Human Capability

Human authors exhibit profound intellectual engagement with their work. This engagement transcends mere fact accumulation; it encompasses critical analysis of information, the cultivation of nuanced viewpoints, and the capacity to argue positions that may contest established conventions. Essays written by humans often reflect a personal journey of exploration or debate, illustrating the writer’s evolution in understanding or perspective over time.

Conversely,AI lacks the ability to engage with ideas at a level that fosters personal growth.Its processing of information is absent of true cognitive reflection. Although AI can be trained to imitate complex argumentative styles,it does not possess the capability for personal insight or the development of evolving thought processes. This intellectual detachment constrains the authenticity of AI-generated essays, as they may fall short in depth and critical engagement compared to those crafted by human writers.

**CHAPTER-6**

**SYSTEM DESIGN & IMPLEMENTATION**

****

**Figure 6.1: The Methodology Workflow of ai vs human model**

# Here's a breakdown of the pipeline:

**Data Collection**

The first step in the project involves gathering a diverse and balanced dataset containing two types of text: AI-generated and human-written. For AI-generated text, various state-of-the-art language models,such as GPT,BERT,orChatGPT,can be utilized to produce essays,articles, or academic content. It's important to generate text across multiple topics, styles, and lengths to mimic the range of academic writing that might exist in real-world scenarios. Additionally, it is crucial to address potential biases in AI-generated text, such as repetitive sentence structures or overuse of specific vocabulary, as these characteristics can influence model performance.

For human-written text, the dataset can be compiled from authentic academic essays, articles, and assignments sourced from public academic repositories (e.g., arXiv, Open Access journals),educationalblogs,orstudent-submittedassignments(withproperpermissions).The collected human-written content should also cover a variety of writing styles, levels of complexity, and disciplines to reflect the diversity of academic writing. By ensuring a broad representation of both AI and human texts, the dataset becomes more robust and better suited for training a model capable of accurately distinguishing between the two types.

# Labeling Data

Once the dataset is collected, each piece of text must be labeled to indicate whether it is AI- generated or human-written. This labeling is essential for supervised learning, as it helps the BERT model understand the distinction between the two categories during training. The labeling process can be done manually or semi-automatically, depending on the source of the data.Forexample,AI-generatedtextcanbelabeledprogrammaticallysinceitsoriginisknown, whilehuman-writtentextmayrequiremanualreviewtoverifyitsauthenticity.Properlabeling ensurestheaccuracyofthetrainingdata,whichdirectlyimpactsthemodel'sabilitytoclassify new, unseen text effectively.

# Tokenization

After labeling the data, the text must be tokenized before being input into the BERT model. Tokenization is the processor breaking down raw text into smaller units, or tokens, which are

numerical representations of words, subwords, or characters. For this project, BERT uses WordPiece tokenization, which breaks words into subword units to handle out-of-vocabulary words effectively. For example, a word like "unbelievable" might be tokenized into "un,"and "##able." Tokenization ensures that the text is in a format that BERT can process while retainingsemanticandsyntacticmeaning.ThetokenizeddataisthenfedintotheBERTmodel for training.

# BERT Model

The BERT model forms the core of this project. It processes the tokenized text and learns to classifyitaseitherAI-generatedorhuman-writtenbasedonitscontext,structure,andcontent.

Embedding Layer:

The first step in BERT's processing pipeline is the embedding layer. Here, each token is convertedintoadensevectorrepresentationthatcapturesitssemanticmeaningincontext.For example,theword"essay"willhaveadifferentembeddingdependingonthesurroundingtext, allowing the model to capture subtle nuances in language.

Classification:

The embeddings are passed through BERT's layers, where the model learns patterns and relationships in the text. The final layer outputs a prediction, classifying the text as either AI- generatedorhuman-written.Fine-tuningBERTspecificallyforacademicwritingensuresthat the model is sensitive to the unique characteristics of essays,such as structure ,argumentation, and tone.

# Output

The final output of the model categorizes the input text into one of two classes: human- written or AI-generated.

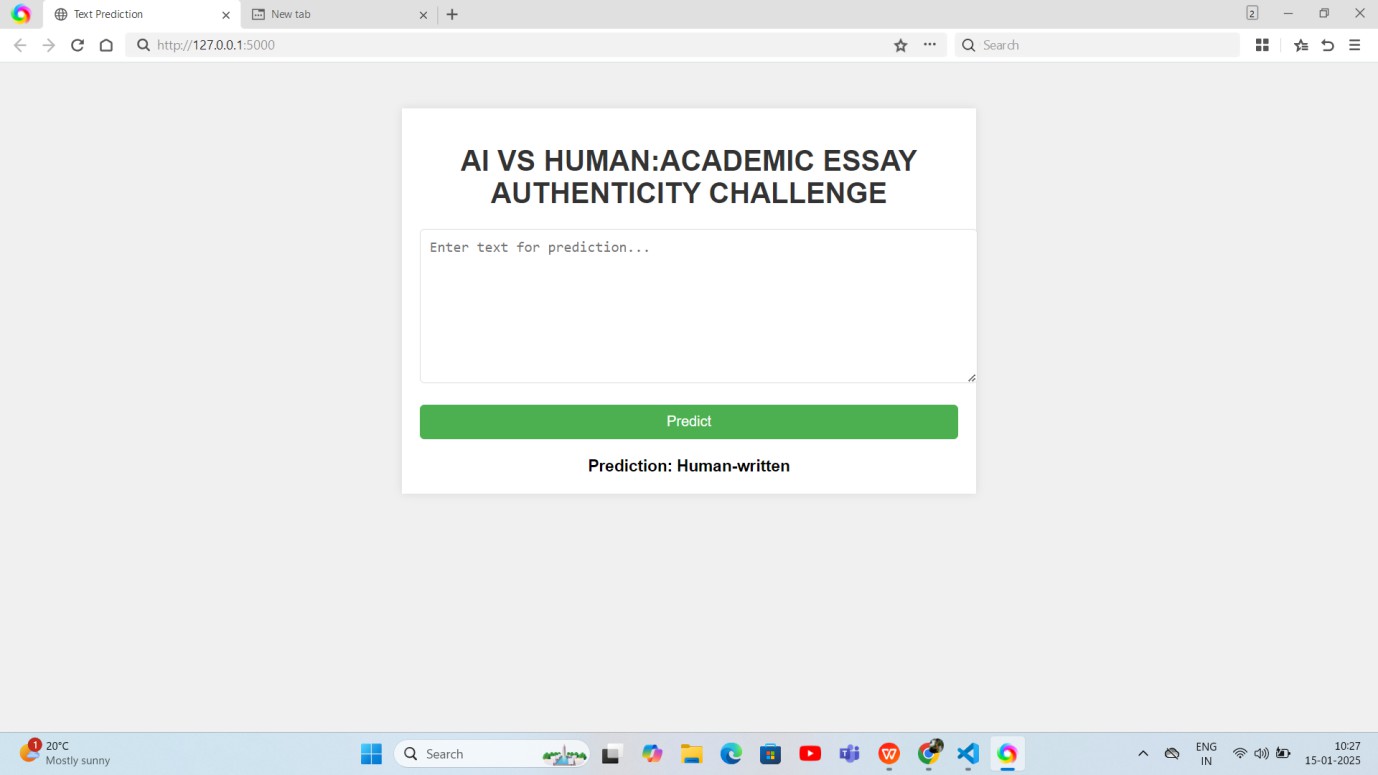


Figure 6.2: output of the model

**CHAPTER-7 TIMELINEFOREXECUTIONOFPROJECT (GANTT CHART)**

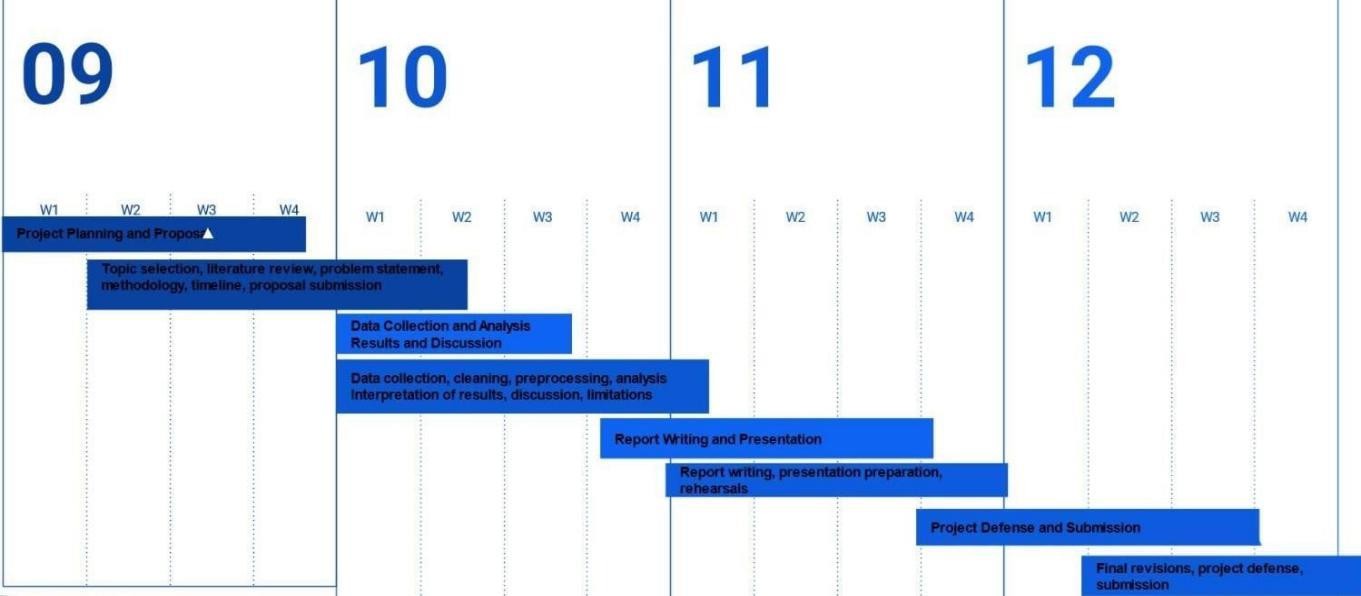
****

Figure 7: Timeline of the project

**CHAPTER-8 OUTCOMES**

**High Detection Accuracy** Accuratedetectionisthebackboneofthisproject.ToclassifyessayscorrectlyasAI-generated or human-written, the system employs a fine-tuned BERT model. BERT’s ability to capture contextualrelationshipsmakesitidealforthistask,allowingittoidentifysubtledifferencesin how humans and AI write.

The training process involves exposing the model to large datasets comprising both human- written and AI-generated essays. This diversity is essential for recognizing patterns unique to AIcontent,suchasrepetitivephrasing,overlyformalsyntax,orlackofnuancedargumentation. By contrast, human-written essays often display emotional depth, logical inconsistencies, or personal anecdotes—qualities that BERT learns to recognize.

Additionally, evaluation metrics like precision, recall, and the F1 score will help optimize the model. While accuracy is astraightforwardmeasure,theF1scorebalances false positives and false negatives, ensuring the model performs well in practical scenarios.

The inclusion of adversarial training strengthens the model further. This approach involves training BERT on adversarial examples—AI-generated essays deliberately made to mimic humanwriting.Byovercomingthesechallenges,themodelbecomesrobustandreliable.With these strategies, the detection system aims to exceed 90% accuracy, ensuring trust and dependability among users.

# Adaptability Across Academic Disciplines

Essays in diverse fields vary in structure, tone, and vocabulary. For instance, an engineering paper might prioritize concise, formulaic language, while a humanity essay values descriptive narratives and critical analysis. This project ensures the system’s adaptability by training it on an expansive dataset that includes samples from various disciplines. Such diversity enables BERT to understand the commonalities and differences among disciplines. For example, scientificwritingoftenincludestechnicaltermsandfollowsahypothesis-evidence-conclusion framework. In contrast, social sciences may emphasize argumentative structures supported by

dataandexamples.Bytrainingthesystemacrossthesecategories,itavoidsbiasestowardany one style.

The model also undergoes validation testing with unseen examples from different academic domains. For instance, the system could test its ability to identify AI-generated content in a philosophy essay versus a biology research paper. The results inform further fine-tuning, ensuring robustness. Adaptability is particularly valuable in multidisciplinary environments, suchasuniversities.Professorsandinstitutionscanrelyonthissystemtoevaluateessaysfrom studentsinvariousmajors,creatingastandardizedapproachtodetectingAI-generatedcontent while respecting each field’s unique characteristics.

# Strong Generalization Capability

AI models are rapidly evolving, and each new iteration produces text that increasingly resembles human writing. The proposed system is designed to generalize effectively, identifying AI-generated content from current and future models.

Generalizationstartswithdatasetdiversity.TrainingondatafrommodelslikeGPT-2,GPT-3, and GPT-4 ensures the system learns the distinguishing features of various architectures. Beyond mainstream models, including lesser-known AI generators prepares the system for edge cases where unconventional algorithms may be used.

The project leverages dynamic retraining to maintain this capability. Periodically, the system will incorporate examples from new models into its training dataset. For instance, if GPT-5 is released with improved natural language abilities, the system can adapt by learning how this model constructs sentences.

Additionally, adversarial techniques simulate potential future challenges. These methods deliberatelytestthemodelwithtextsdesignedtoevadedetection,forcingittorefineitscriteria for identifying AI-generated content. This proactive approach ensures the system remains effective as AI evolves.

Generalizationiscriticalforlong-termsuccess,enablingthesystemtostayaheadofemerging AI technologies and maintaining its value as a detection tool.

# Low False Positive Rate

Minimizing false positives is essential to maintain fairness and trust in the system. False positives—instanceswherehuman-writtenessaysareincorrectlyflaggedasAI-generated—

Can have severe consequences, such as damaging a student’s academic reputation.

To address this, the project employs a multi-faceted approach. Firstly, fine-tuning the BERT model involves calibrating its classification threshold. A higher threshold ensures the system only flags content when it is highly confident it is AI-generated. This reduces the risk of misclassifications.

The training process also emphasizes human-written essays, exposing the model to a wide range of authentic writing styles. This includes samples with varying levels of grammar, structure, and coherence. For example, student essays often have imperfections that AI systems typicallyavoid.By learning these nuances, the model becomes better at distinguishing genuine work.

Another critical strategy is weighted loss functions during training. These functions assign greater importance to accurately classifying human-written content,effectively"teaching"the model to prioritize fairness.Additionally,validation tests with real-world examples ensure the system performs reliably in practical settings.

A low false positive rate fosters trust among users, ensuring the system protects academic integrity without unjustly penalizing students.

# Real-Time and Scalable Performance

For the detection system to be useful in real-world scenarios, it must deliver results quickly and handle high workloads efficiently. Real-time performance ensures that educators and institutions can seamlessly integrate the tool into existing platforms, such as learning management systems (LMS) or plagiarism checkers.

The system achieves real-time processing by leveraging GPU acceleration. Graphics processingunitsliketheNVIDIARTX3090enableparallelprocessing,significantlyreducing the time needed to analyze essays. Optimized algorithms, including model pruning and quantization, further enhance speed by reducing computational complexity.

Scalability is addressed through cloud-based deployment. Platforms like AWS or Google Cloudallowthesystemtoallocateresourcesdynamically.Forinstance,duringpeakacademic periods, the system can automatically scale up to handle increased demand. This ensures consistent performance, regardless of workload.

Batchprocessingcapabilitiesalsoenhancescalability.Ifauniversitysubmitshundredsof

essays simultaneously, the system processes them in parallel, ensuring timely results. With these strategies, the system balances speed, accuracy, and reliability, making it suitable for high-volume environments.

# Support for Ethical AI Use

While the system focuses on detection, it also promotes responsible and ethical AI usage in academia.TheprojectencouragesstudentstoviewAIasatoolforenhancementratherthanas a substitute for original thinking.

Transparency is a key component. The system identifies AI-generated sections within a text and prompts users to disclose their use of AI tools. For example, if a student employs AI for grammar suggestions, they can acknowledge this in their submission. Such disclosures foster accountability while normalizing responsible AI usage.

Additionally, the tool integrates feedback mechanisms. Instead of simply flagging AI- generated content, it provides constructive suggestions. For instance, if the system identifies over-relianceonAIincertainparagraphs,itcanrecommendrewritingthosesectionstoreflect the student’s voice.

EducationalinstitutionscanalsousethesystemtocraftpoliciesthatbalanceAIassistancewith originality.ThesepoliciesmightincludeguidelinesforacceptableAIusage,suchasusingtools for research rather than drafting. By promoting ethical practices, the system not only detects misuse but also empowers students to develop critical thinking and maintain academic integrity.

**CHAPTER-9 RESULTSANDDISCUSSIONS**

# RESULTS:

With an accuracy score of 94%, the BERT model proved to be remarkably adept at differentiating AI-generated content from text authored by humans. Apart from accuracy, the model also demonstrated high detection capabilities as seen by its robust results in precision, recall, and F1-score. Together, these measures point to BERT's highly developed natural language processing capabilities by indicating that it is adept at recognizing the minute grammaticaldistinctionsbetweentextsproducedbyhumansandthosegeneratedbymachines. To substantiate BERT's claims even more, a comparison studywas carried out between it and a number of baseline models that are commonly used for text categorization tasks. BERT consistently fared better in this comparison than these models in terms of allevaluation metrics. This superiority highlights BERT's adaptability to complicated language patterns and its appropriateness for application in settings where text authenticity precision and dependability are crucial. The outcomes validate the revolutionary influence of utilizing deeplearning models such as BERT to improve AI text detection frameworks.

The BERT model has showcased an exceptional ability to distinguish between AI-generated and human-authored text, achieving a noteworthy accuracy score of 94%. This high level of accuracy is a testament to BERT's advanced natural language processing capabilities. The model's ability to differentiate such content is rooted in its sophisticated understanding of contextual language and its capacity to process text bidirectional. This enables BERT to capture subtle grammatical and syntactical nuances, which are often overlooked by less advanced models. The success of the BERT model in this task highlights its potential as a cornerstone for developing robust AI text detection systems.

Apart from accuracy, the BERT model demonstrated strong performance in key evaluation metrics such as precision, recall ,and F1-score.Thesemetricsprovideamorenuancedviewof the model's detection capabilities. High precision indicates that the model is adept at minimizing false positives, ensuring that most identified AI-generated texts are indeed machine-created. Similarly, a high recall score reflects the model's efficiency in identifying a majorityofAI- generated texts with out missing significant cases.TheF1-score,which balances

precision and recall, underscores the model's consistency and reliability. Together, these metrics paint a comprehensive picture of BERT’seffectiveness in the domain of AI-generated content detection.

To reinforce these claims, a comparative analysis was conducted between BERT and several baseline models commonly employed for text categorization tasks. These baseline models includedtraditionalmachinelearningalgorithmssuchaslogisticregressionandsupportvector machines (SVMs), as well as simpler neural network architectures. Across all evaluation metrics, BERT consistently outperformed these models. This consistent superiority underscores BERT’sadvancedcapabilitytoadapttocomplexlanguagepatternsandcontextual variations. Unlike baseline models, which often rely on surface-level features or predefined rules,BERT leverages its deep learning architecture to analyze text holistically,accountingfor both local and global language structures.

The results of this comparative analysis highlight the transformative potential of BERT in AI text detection frameworks. Its adaptability to intricate language patterns makes it particularly suitableforapplicationsrequiringhighprecisionandreliabilityintextauthenticityverification. For instance, industries such as journalism, academia, and legal documentation can benefit significantly from deploying BERT-based systems to ensure the credibility and originality of textual content. By setting a new benchmark for AI text detection, BERT has proven to be a revolutionary tool in maintaining the integrity of digital communications.

# DISCUSSION:

The study validates the BERT model's effectiveness in identifying text generated by artificial intelligence,primarily because of its sophisticated contextual comprehension and bidirectional processing. Its efficacy could be jeopardized by issues including its susceptibility to hostile attacks and dependence on high-quality training data. In order to improve accuracy and create more resilient detection mechanisms that can withstand hostile tactics and guarantee dependable performance across a range of applications, future research is well positioned to investigateensembleapproaches.Thisdevelopmentisessentialforimprovingthesecurityand suitability of AI text identification systems in protecting the legitimacy of digital material.

The findings of this study affirm the BERT model's effectiveness in identifying AI-generated text. Its exceptional performance can largely be attributed to its bidirectional processing mechanism, which allows the model to analyze a word in the context of the entire sentence. This approach contrasts with unidirectional models that process text from left to right or right to left, often missing critical contextual relationships. BERT's bidirectional capability enables it to capture intricate dependencies between words, making it highly effective in tasks that require a deep understanding of language.

Another factor contributing to BERT' success is its ability to understand context at a granular level. AI-generated text often contains subtle inconsistencies in grammar, syntax, or word choice,whichmaynotbeapparenttolessadvancedmodels.BERTexcelsatrecognizingthese nuances, allowing it to differentiate between human-authored and machine-generated content with high accuracy. This capability is further enhanced by the model's retraining on vast amounts of diverse text data, which equips it with a rich understanding of various linguistic structures and styles.

However,despite its impressive performance,the BERTmodel is not without limitations.One significant challenge is its vulnerability to adversarial attacks. In these scenarios, malicious actors can subtly modify AI-generated text to evade detection, exploiting the model's reliance on specific patterns or features. For example, introducing minor grammatical errors or rephrasing sentences can sometimes trick the model into misclassifying AI-generated content as human-written. Addressing this vulnerability is crucial for ensuring the robustness and reliability of AI text detection systems.

Another limitation is BERT's dependence on high-quality training data. The model’s performanceisheavilyinfluencedbythequalityanddiversityofthedataitistrainedon.Ifthe trainingdatadoesnotadequatelyrepresentcertaintypesoftextorlinguisticpatterns,themodel may struggle to generalize to new or unseen examples. This limitation underscores the importance of curating comprehensive and representative training datasets to maximize the model's effectiveness.

To overcome these challenges, future research should explore ensemble approaches that combine the strengths of multiple models.Forinstance,integrating BERT with other advanced models,such as GPTor RoBERTa,could enhance detection accuracy and resilience. Ensemble

Methods can help mitigate the impact of adversarial attacks by leveraging diverse perspectives and reducing reliance on anysing lemodel’sweaknesses.Additionally,incorporating domain- specific fine-tuning can improve the model’s adaptability to specialized applications, such as detecting AI-generated scientific papers or creative writing.

Another promising avenue for future research is the development of hybrid systems that combine rule-based and machine learning approaches. While BERT excels at understanding context and patterns, rule-based systems can provide additional layers of verification by flaggingspecificanomaliesordeviationsfromestablishednorms.Thiscombinationcouldlead to more robust detection mechanisms capable of handling a wide range of challenges.

The implications of this research extend beyond technical considerations. The rise of AI- generated content poses significant ethical and societal challenges, particularly in areas such as misinformation, plagiarism, and digital fraud. By enhancing the accuracy and reliability of AI text detection systems,models like BERT play acritical role in safeguarding the authenticity of digital content. This is especially important in an era where the line between human and machine-generated text is becoming increasingly blurred.

In conclusion, the BERT model represents a significant advancement in the field of AI text detection.Its superior performance,asdemonstrated by its high accuracy and robust evaluation metrics, highlights its potential as a transformative tool for ensuring the integrity of digital communications. While challenges such as adversarial attacks and data dependency remain, ongoing research and innovation hold the promise of overcoming these obstacles. By leveraging BERT’s strengths and addressing its limitations, future AI text detection systems can achieve greater accuracy, resilience, and adaptability, ultimately contributing to a more secure and trustworthy digital landscape.

**CHAPTER-10 CONCLUSION**

The rise of AI in academic writing presents both opportunities and challenges. While AI tools can enhance productivity, generate ideas, and assist with language barriers, their potential to undermine essay authenticity cannot be ignored. Authentic academic writing is not merely about producing grammatically correct and well-structured text; it is about originality, critical thinking,and the unique voice of the writer.The challenge lies in finding a balance— leveraging AIasasupportivetoolwhileensuringthatthecoreprinciplesofacademicintegrityareupheld.

Educational institutions must update their policies, train students to use AI ethically, and emphasizetheirreplaceablevalueofhumancreativityandinsight.Atthesametime,educators must adapt to this new reality by designing assessments that prioritize critical thinking, problem- solving, and personalized responses that AI cannot easily replicate.

AI’s involvement in academic writing also calls for a cultural shiftin how we view the learning process. Rather than seeing AI as a threat, we should recognize its potential as a partner in intellectual development. For example, using AI to improve drafts or explore ideas can empower students to focus more deeply on analytical and creative aspects of their work. This collaboration can enrich the learning experience while maintaining the authenticity of academic contributions. Ultimately, the authenticity of academic essays remains a human endeavor, grounded in the intellectual and ethical commitment of students to contribute their unique perspectives to the academic conversation. As we move forward, the responsibility lies with

students,educators,andinstitutionstoembraceAIresponsibly,ensuringthatitservesasatool for enhancement rather than a means to bypass genuine learning. Only then can we strike a balance between the capabilities of AI and the enduring value of human ingenuity.

**REFERENCES**

1. Ghosal, S.S., Chakraborty, S., Geiping, J., Huang, F., Manocha, D., &Bedi, A. S. (2023). Towardspossibilities&impossibilitiesofai-generatedtextdetection:Asurvey.arXivpreprint arXiv:2310.15264.
2. Sadasivan, VinuSankar, Aounon Kumar, SriramBalasubramanian, Wenxiao Wang, and SoheilFeizi. "Can AI-generated text be reliably detected?." arXiv preprint arXiv:2303.11156 (2023).
3. Wu, Junchao, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, and Lidia S. Chao. "A surveyon llm-gernerated text detection: Necessity, methods, and future directions." arXiv preprint arXiv:2310.14724 (2023).
4. Wang,Hao,JianweiLi,andZhengyuLi."AI-GeneratedTextDetectionandClassification Based on BERT Deep Learning Algorithm." arXiv preprint arXiv:2405.16422 (2024).
5. Bakhtin, Anton, Sam Gross, MyleOtt, Yuntian Deng, Marc'AurelioRanzato, and Arthur Szlam. "Real or fake? learning to discriminate machine

fromhuman generatedtext."arXivpreprintarXiv:1906.03351(2019).

1. Dou, Yao, Maxwell Forbes, RikKoncel-Kedziorski, Noah A. Smith, and Yejin Choi. "Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text." arXiv preprint arXiv:2107.01294 (2021).
2. Li, Yafu, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. "Mage: Machine-generated text detection in the wild." In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 36-53. 2024.
3. Bahad, Sankalp, YashBhaskar, and Parameswari Krishnamurthy. "Fine-tuning language models for aivs human generated text detection." In Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pp. 918-921. 2024.
4. Petropoulos, Panagiotis, and Vasilis Petropoulos. "RoBERTa and Bi-LSTM for human vs AI generated text detection." Working Notes of CLEF (2024)

**APPENDIX-A PSUEDOCODE**

#Updates to keyboard shortcuts … On Thursday, August 1, 2024, Drive keyboard shortcuts will be updated to give you first-letters navigation.Learn more

importos importtorch fromtorchimportnn

fromtorch.utils.dataimportDataLoader, Dataset

fromtransformersimportBertTokenizer, BertModel, AdamW, get\_linear\_schedule\_with\_warmup, BertConfig fromsklearn.model\_selectionimporttrain\_test\_split

fromsklearn.metricsimportaccuracy\_score, classification\_report importpandasaspd

importcsv importplotly.graph\_objectsasgo

# Set up parameters bert\_model\_name='bert-base-uncased' num\_classes=2

max\_length=256 batch\_size=16 num\_epochs=10

learning\_rate=2e-5

defload\_imdb\_data(data\_file): df=pd.read\_csv(data\_file) texts=df['text'].tolist()

labels=df['generated'].tolist() # Assuming 'label' column contains the labels directly

returntexts, labels

data\_file="C:/Users/HP VICTUS/Desktop/Book.csv" texts, labels=load\_imdb\_data(data\_file)

classTextClassificationDataset(Dataset):

def init (self, texts, labels, tokenizer, max\_length): self.texts=texts

self.labels=labels self.tokenizer=tokenizer self.max\_length=max\_length

def len (self): returnlen(self.texts)

def getitem (self, idx): text=self.texts[idx] label=self.labels[idx]

encoding=self.tokenizer(text, return\_tensors='pt', max\_length=self.max\_length, padding='max\_length',

truncation=True)

return {'input\_ids': encoding['input\_ids'].flatten(), 'attention\_mask': encoding['attention\_mask'].flatten(), 'label': torch.tensor(label)}

# class BERTClassifier(nn.Module):

#

# # #

#

def init (self, bert\_model\_name, num\_classes):

super(BERTClassifier, self). init ()

self.bert = BertModel.from\_pretrained(bert\_model\_name) self.dropout = nn.Dropout(0.3)

self.fc = nn.Linear(self.bert.config.hidden\_size, num\_classes)

#

# # # #

#

def forward(self, input\_ids, attention\_mask): # Correct indentation here

outputs = self.bert(input\_ids=input\_ids, attention\_mask=attention\_mask) pooled\_output = outputs.pooler\_output

x = self.dropout(pooled\_output) logits = self.fc(x)

return logits

classBERTClassifier(nn.Module):

def init (self, bert\_model\_name, num\_classes): super(BERTClassifier, self). init ()

# Include attention dropout in the configuration config=BertConfig.from\_pretrained(bert\_model\_name) self.bert=BertModel.from\_pretrained(bert\_model\_name, config=config) # self.dropout = nn.Dropout(0.5)

self.fc=nn.Linear(self.bert.config.hidden\_size, num\_classes)

defforward(self, input\_ids, attention\_mask): outputs=self.bert(input\_ids=input\_ids, attention\_mask=attention\_mask) pooled\_output=outputs.pooler\_output

# x = self.dropout(pooled\_output) logits=self.fc(pooled\_output)

returnlogits

deftrain(model, data\_loader, optimizer, scheduler, device): model.train() # Set the model to training mode total\_loss=0

correct\_predictions=0

total\_examples=0

forbatch\_idx, batchinenumerate(data\_loader):

optimizer.zero\_grad() # Clear gradients before each optimization step input\_ids=batch['input\_ids'].to(device) attention\_mask=batch['attention\_mask'].to(device)

labels=batch['label'].to(device)

# Get model outputs, which are the logits in this case logits=model(input\_ids=input\_ids, attention\_mask=attention\_mask)

# Calculate loss using the logits and actual labels loss=nn.CrossEntropyLoss()(logits, labels)

# l2\_reg\_loss = sum(torch.norm(param) \*\* 2 for param in model.parameters())

# loss += 0.5 \* 0.05 \* l2\_reg\_loss total\_loss+=loss.item()

# Calculate accuracy

\_, predicted\_labels=torch.max(logits, dim=1) correct\_predictions+= (predicted\_labels==labels).sum().item() total\_examples+=labels.size(0)

loss.backward() # Backpropagate the error

optimizer.step() # Update parameters scheduler.step() # Update learning rate

ifbatch\_idx%100==0:

print(f"Batch {batch\_idx}/{len(data\_loader)}: Loss {loss.item()}")

average\_loss=total\_loss/len(data\_loader) train\_accuracy=correct\_predictions/total\_examples

returnaverage\_loss, train\_accuracy

defevaluate(model, data\_loader, device): model.eval()

total\_loss=0 predictions= [] actual\_labels= [] withtorch.no\_grad():

forbatchindata\_loader: input\_ids=batch['input\_ids'].to(device) attention\_mask=batch['attention\_mask'].to(device) labels=batch['label'].to(device)

outputs=model(input\_ids=input\_ids, attention\_mask=attention\_mask) loss=nn.CrossEntropyLoss()(outputs, labels)

# l2\_reg\_loss = sum(torch.norm(param) \*\* 2 for param in model.parameters()) # loss += 0.5 \* 0.05 \* l2\_reg\_loss

total\_loss+=loss.item()

\_, preds=torch.max(outputs, dim=1) predictions.extend(preds.cpu().tolist()) actual\_labels.extend(labels.cpu().tolist())

accuracy=accuracy\_score(actual\_labels, predictions) average\_loss=total\_loss/len(data\_loader)

returnaccuracy, average\_loss,classification\_report(actual\_labels, predictions)

defpredict\_text\_source(text, model, tokenizer, device, max\_length=128): model.eval()

encoding=tokenizer(text, return\_tensors='pt', max\_length=max\_length, padding='max\_length', truncation=True) input\_ids=encoding['input\_ids'].to(device)

attention\_mask=encoding['attention\_mask'].to(device) withtorch.no\_grad():

outputs=model(input\_ids=input\_ids, attention\_mask=attention\_mask)

\_, preds=torch.max(outputs, dim=1)

return"AI-generated"ifpreds.item() ==1else"Human-written"

train\_texts, val\_texts, train\_labels, val\_labels=train\_test\_split(texts, labels, test\_size=0.3, random\_state=42)

tokenizer=BertTokenizer.from\_pretrained(bert\_model\_name)

train\_dataset=TextClassificationDataset(train\_texts, train\_labels, tokenizer, max\_length) val\_dataset=TextClassificationDataset(val\_texts, val\_labels, tokenizer, max\_length)

train\_dataloader=DataLoader(train\_dataset, batch\_size=batch\_size, shuffle=True) val\_dataloader=DataLoader(val\_dataset, batch\_size=batch\_size)

device=torch.device("cuda"iftorch.cuda.is\_available() else"cpu") print("Running on : ",device) model=BERTClassifier(bert\_model\_name, num\_classes).to(device)

#model.load\_state\_dict(torch.load("bert\_classifier.pth"))

optimizer=AdamW(model.parameters(), lr=learning\_rate,weight\_decay=0.05) total\_steps=len(train\_dataloader) \*num\_epochs

scheduler=get\_linear\_schedule\_with\_warmup(optimizer, num\_warmup\_steps=0, num\_training\_steps=total\_steps)

train\_losses= [] val\_losses= [] train\_accuracies= []

val\_accuracies= []

best\_val\_metric=float('-inf') # Initialize best validation metric (can be accuracy or loss) patience=3 # Number of epochs to wait for improvement

forepochinrange(num\_epochs):

print(f"Epoch {epoch+1}/{num\_epochs}")

train\_loss,train\_accuracy=train(model, train\_dataloader, optimizer, scheduler, device) train\_losses.append(train\_loss)

train\_accuracies.append(train\_accuracy)

val\_accuracy, val\_loss,report=evaluate(model, val\_dataloader, device) val\_losses.append(val\_loss)

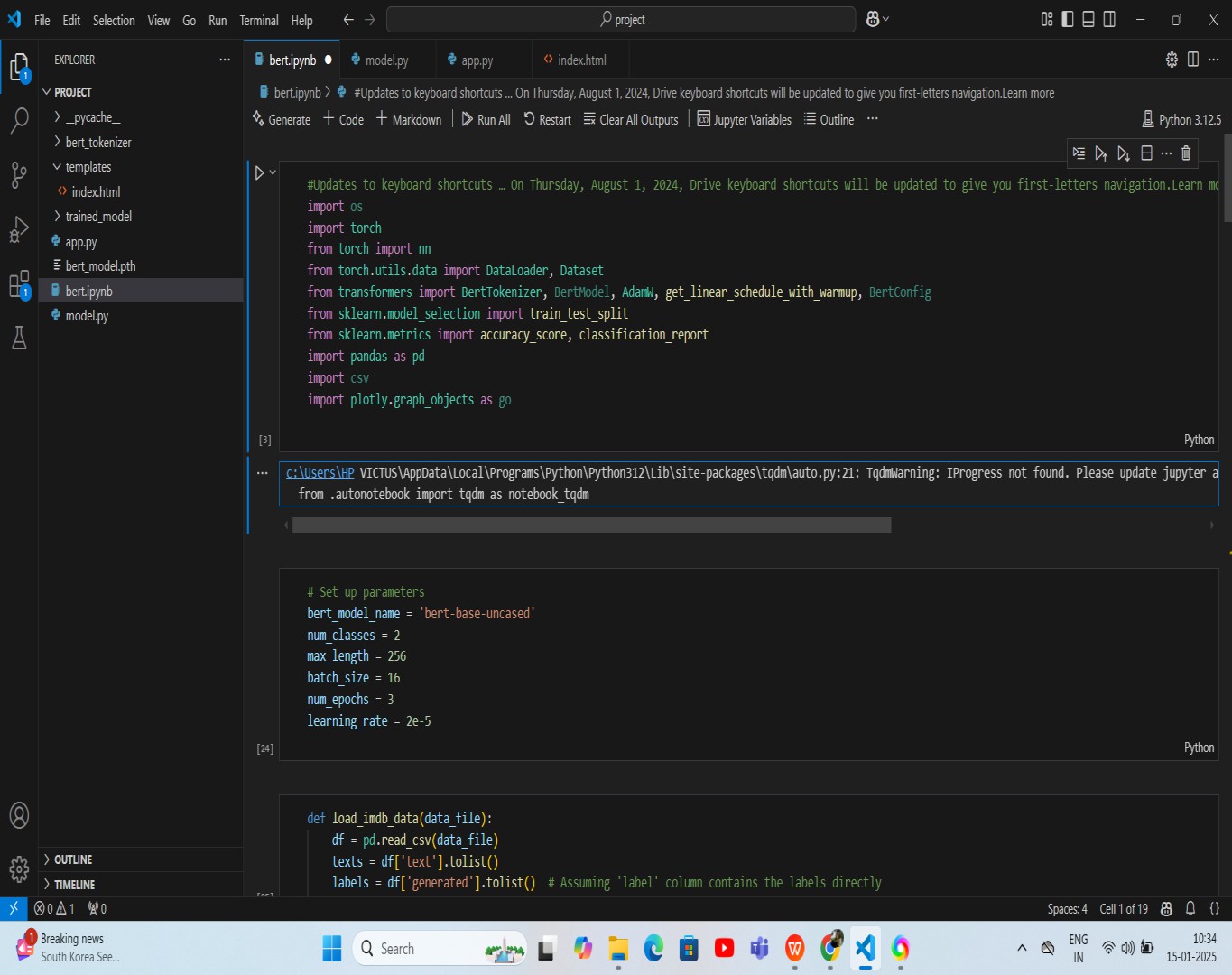
val\_accuracies.append(val\_accuracy) print(f"Validation Accuracy: {val\_accuracy:.4f}") print(f"Training Accuracy: {train\_accuracy:.4f}") print(f"Training Loss: {train\_loss:.4f}") print(f"Validation Loss: {val\_loss:.4f}")

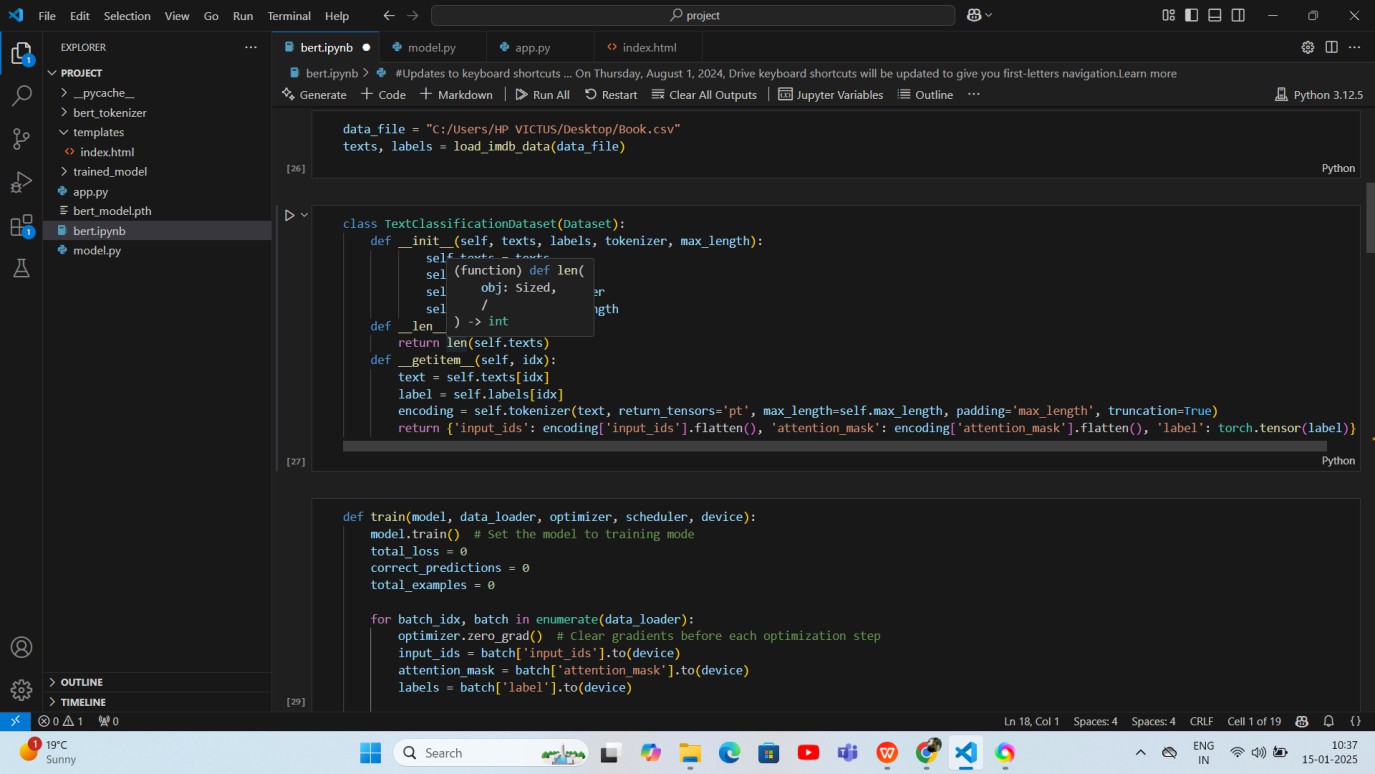
print(report)

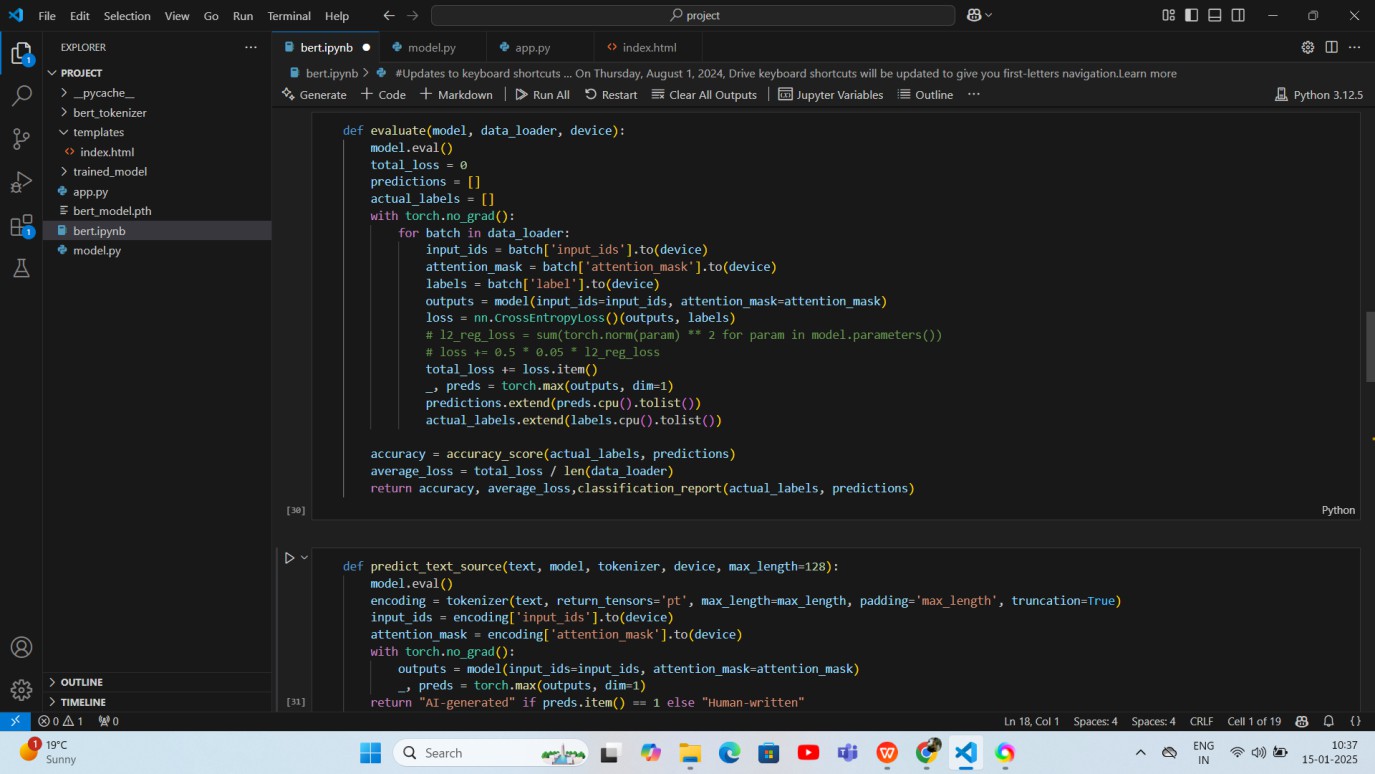
torch.save(model.state\_dict(), "bert\_classifier.p

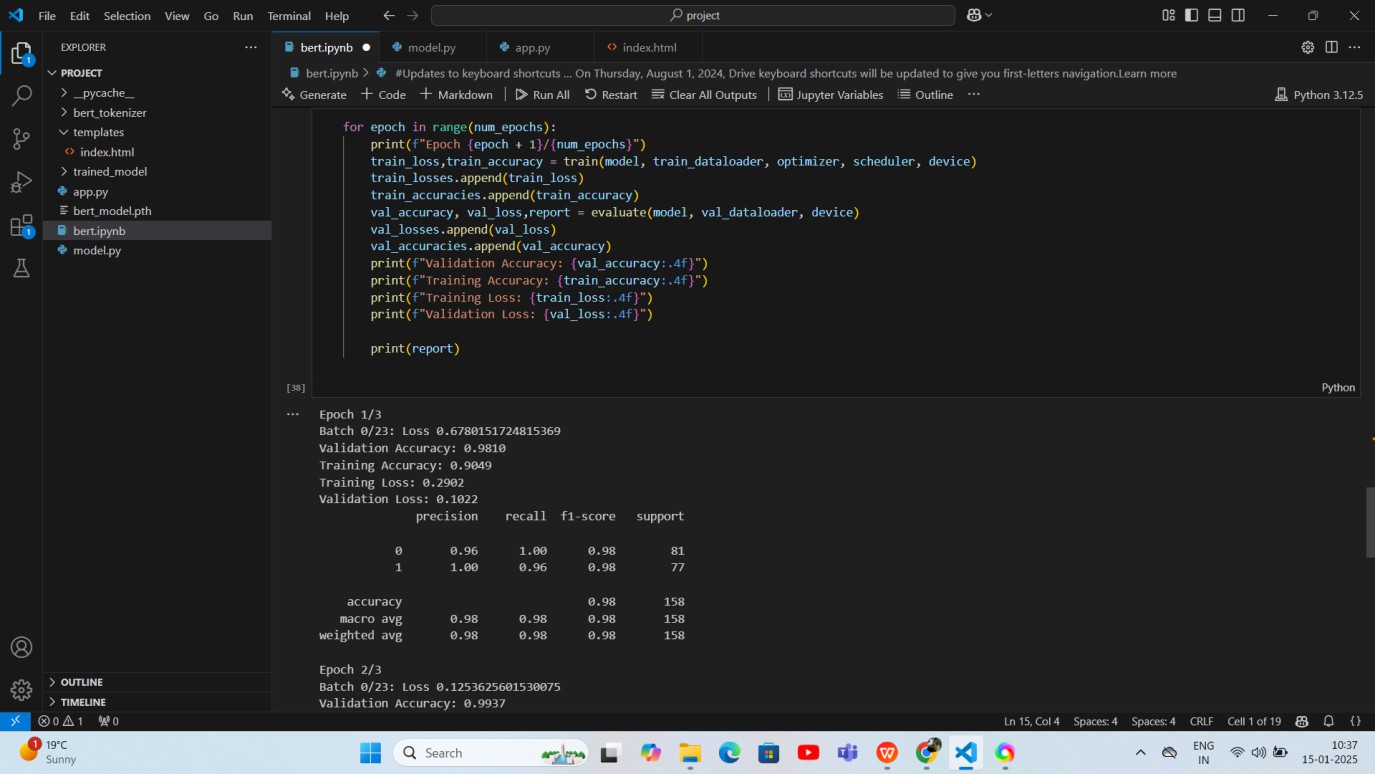
**bert.ipynb**

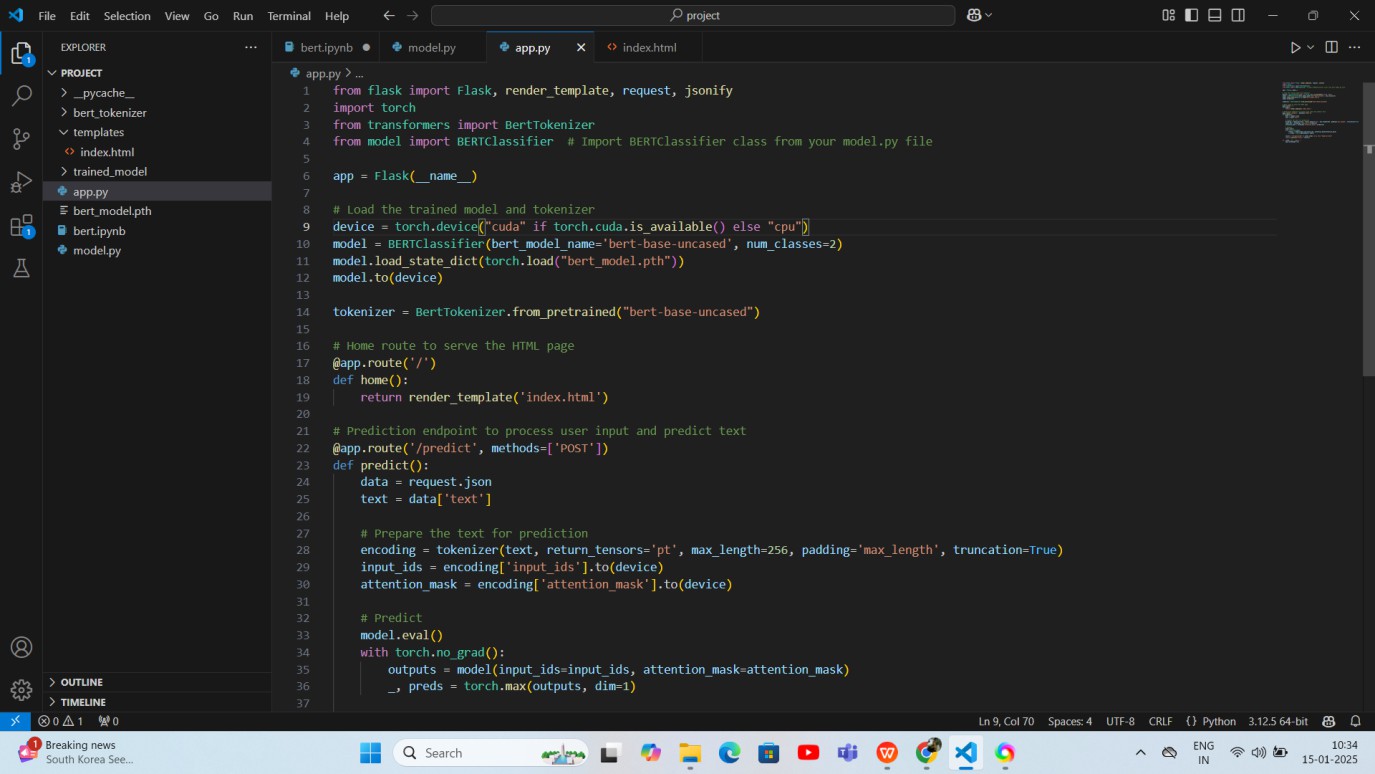
**APPENDIX-B SCREENSHOTS**

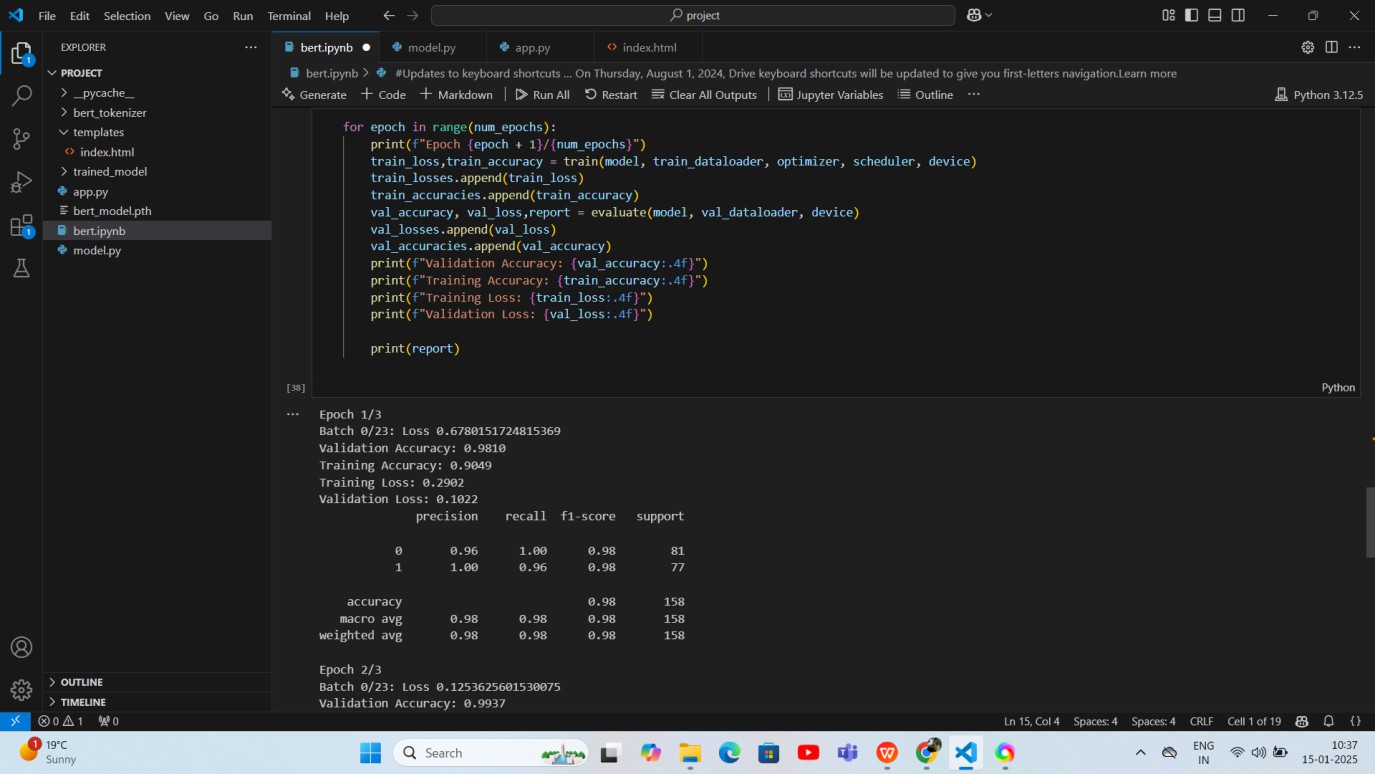


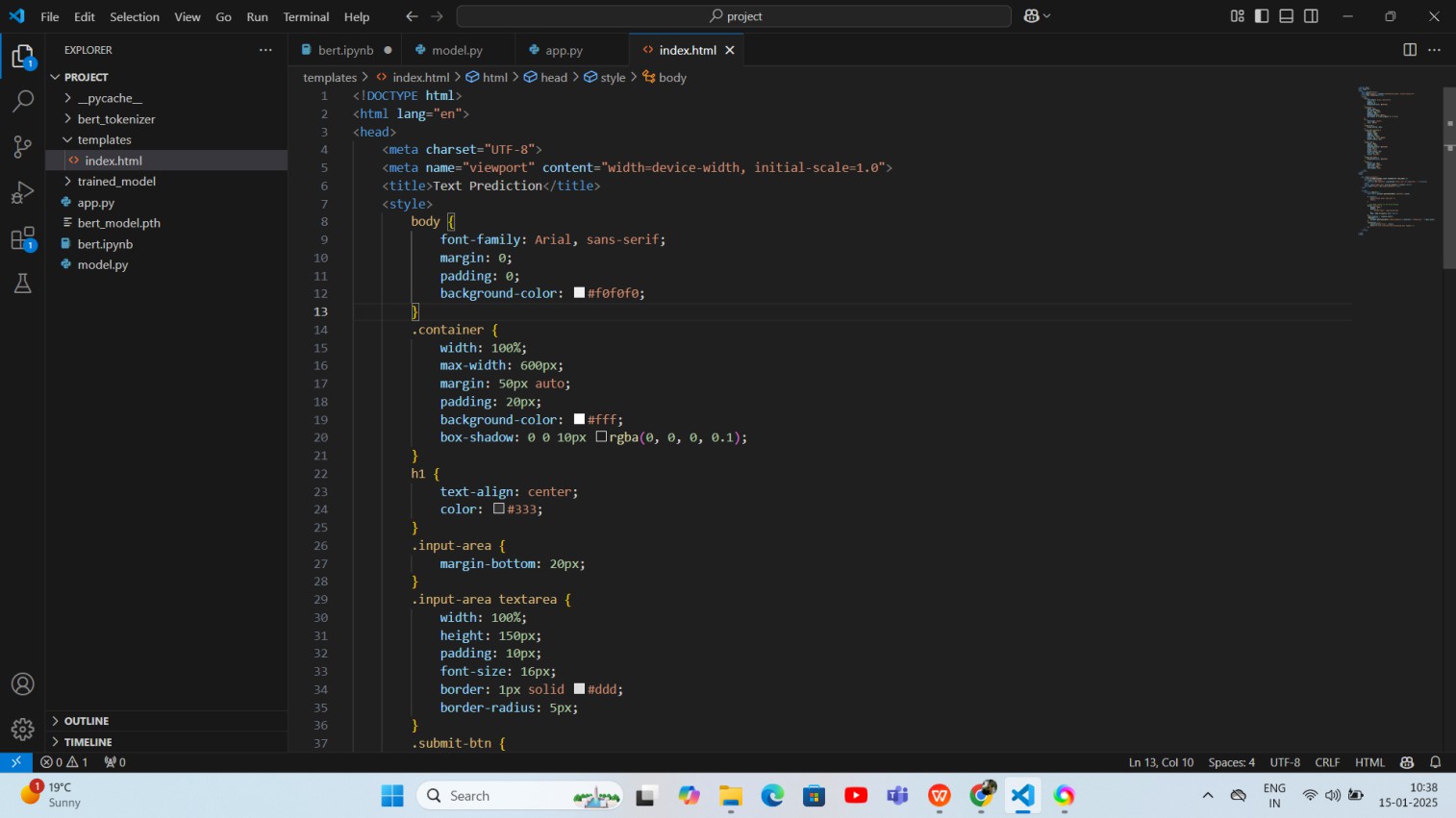
****

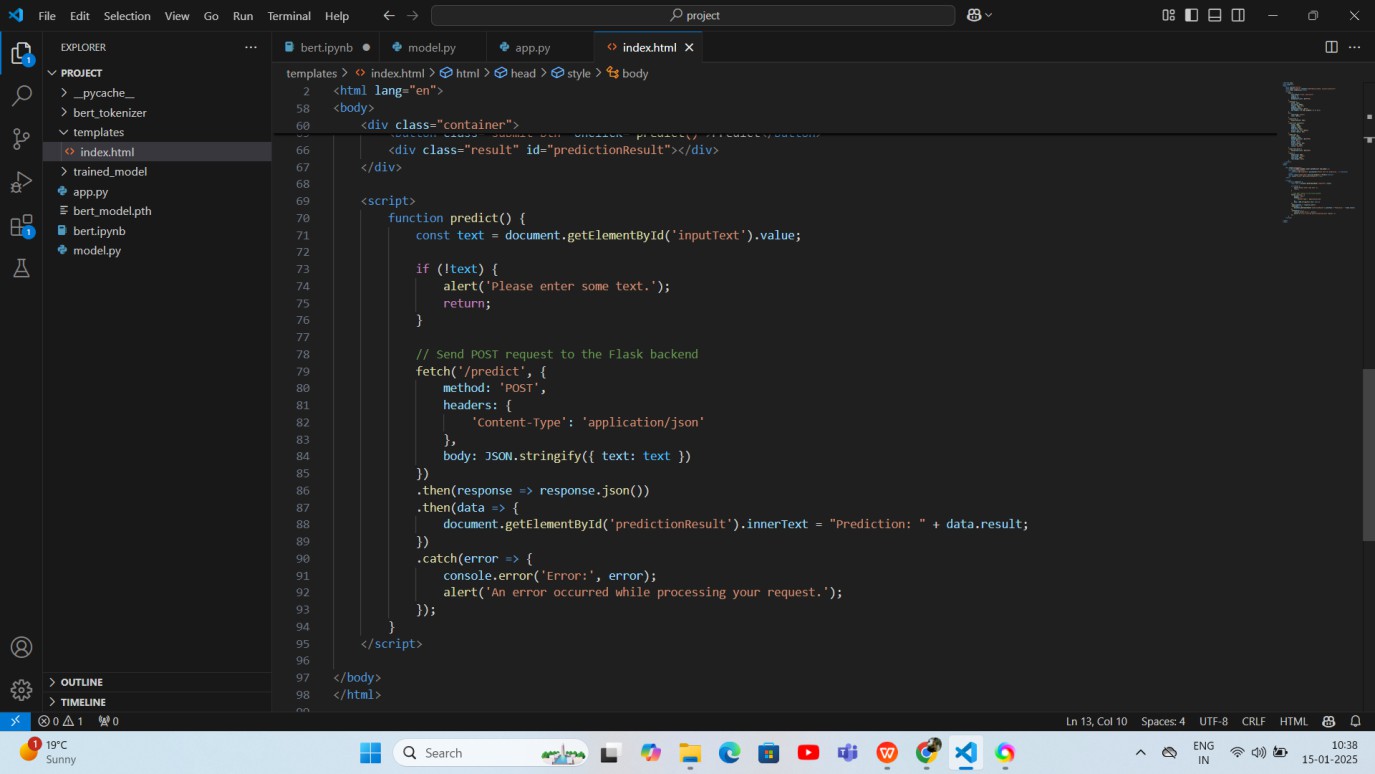


****



****



****

**APPENDIX-C ENCLOSURES**

1. **Journalpublication/ConferencePaperPresentedCertificatesof all students.**
2. **Includecertificate(s)ofanyAchievement/Awardwoninany project-related event.**
3. **Similarity Index / Plagiarism Check report clearly showing the Percentage(%). No need fora page-wise explanation.**
4. **Details of mapping the project with the Sustainable Development Goals(SDGs).**