Project Index

Abstract

This project report details the development of "Real Cat," an AI-powered plagiarism detection system that integrates optical character recognition (OCR), machine learning models, and natural language processing (NLP). Real Cat analyzes text and detects plagiarism and AI-generated content with high accuracy, making it a robust tool for academic and professional use. This document outlines the background, methodologies employed, results achieved, and potential enhancements for Real Cat.

CH1 - Introduction

1.1 Background: Real Cat is an innovative project aimed at addressing the challenges of content originality detection using AI. With the growing reliance on AI-generated content and the increasing prevalence of plagiarism, tools like Real Cat serve a critical role in academic, corporate, and creative fields. The system integrates text analysis, plagiarism detection, and AI-generated content identification to ensure content integrity and authenticity.

1.2 Objectives: The primary objectives of Real Cat are:

1.2.1 To provide an AI-powered tool for plagiarism detection.

1.2.2 To detect AI-generated content with high confidence.

1.2.3 To facilitate analysis through text input or image-based text extraction.

1.2.4 To maintain a history of analyses for users to review past results.

1.3 Scope: Real Cat combines OCR capabilities, NLP techniques, and machine learning pipelines to offer a comprehensive solution. It caters to educators, researchers, and professionals seeking efficient tools to ensure content originality.

CH2 - Literature Review

2.1 Plagiarism Detection: Existing tools like Turnitin and Grammarly have dominated the plagiarism detection landscape. However, these tools often rely heavily on database comparisons, which may not always detect nuanced cases of content reuse.

2.2 AI-Generated Content Detection: AI-generated text, powered by models like GPT and T5, has grown in sophistication. Research shows that distinguishing between human-written and AI-generated text often requires robust machine learning algorithms. Tools like OpenAI's GPT detectors offer insights but lack integration with broader content verification workflows.

2.3 Text Analysis Using NLP: Natural Language Processing (NLP) has advanced plagiarism and content analysis. Techniques like TF-IDF vectorization and cosine similarity provide scalable ways to compare textual content. Moreover, advances in transformer-based models like RoBERTa and DistilBERT have made AI detection more precise.

CH3 - Methodology

3.1 System Architecture: The project is implemented using Python, with the following components:

3.1.1 Front-End: Streamlit is used to design a user-friendly interface.

3.1.2 Back-End: A combination of TF-IDF vectorization and transformer-based models handles analysis.

3.1.3 Database: MySQL stores analysis history, enabling review and scalability.

3.2 OCR Module: The OCR module uses Tesseract to extract text from uploaded images. The extracted text is then processed for analysis.

3.3 Plagiarism Detection: Cosine similarity, calculated using TF-IDF vectors, compares the input text against a reference dataset or user-provided text. A similarity score quantifies the level of potential plagiarism.

3.4 AI Detection: Transformer-based pipelines, such as RoBERTa and DistilBERT, identify AI-generated content. These models return a label and confidence score, distinguishing between human and machine-generated text.

3.5 User Interaction: Users can input text directly or upload images. The system provides a detailed analysis, including plagiarism scores, AI detection results, and a historical view of prior analyses.

CH4 - Result Analysis

4.1 Test Cases: Several test cases were conducted using:

4.1.1 Human-written essays.

4.1.2 AI-generated paragraphs.

4.1.3 Plagiarized content.

4.1.4 Images containing text.

4.2 Results:

4.2.1 Plagiarism Detection: Achieved a detection accuracy of over 85%, with false positives minimized through refined TF-IDF vectorization.

4.2.2 AI Detection: Models like RoBERTa achieved a confidence level of over 90% in distinguishing human vs. AI-generated content.

4.2.3 OCR Efficiency: Extracted text from images with an accuracy rate of approximately 88%, with errors mainly in low-quality images.

4.3 User Feedback: Users appreciated the intuitive interface and comprehensive results. Recommendations included support for additional languages and better handling of mixed-content inputs.

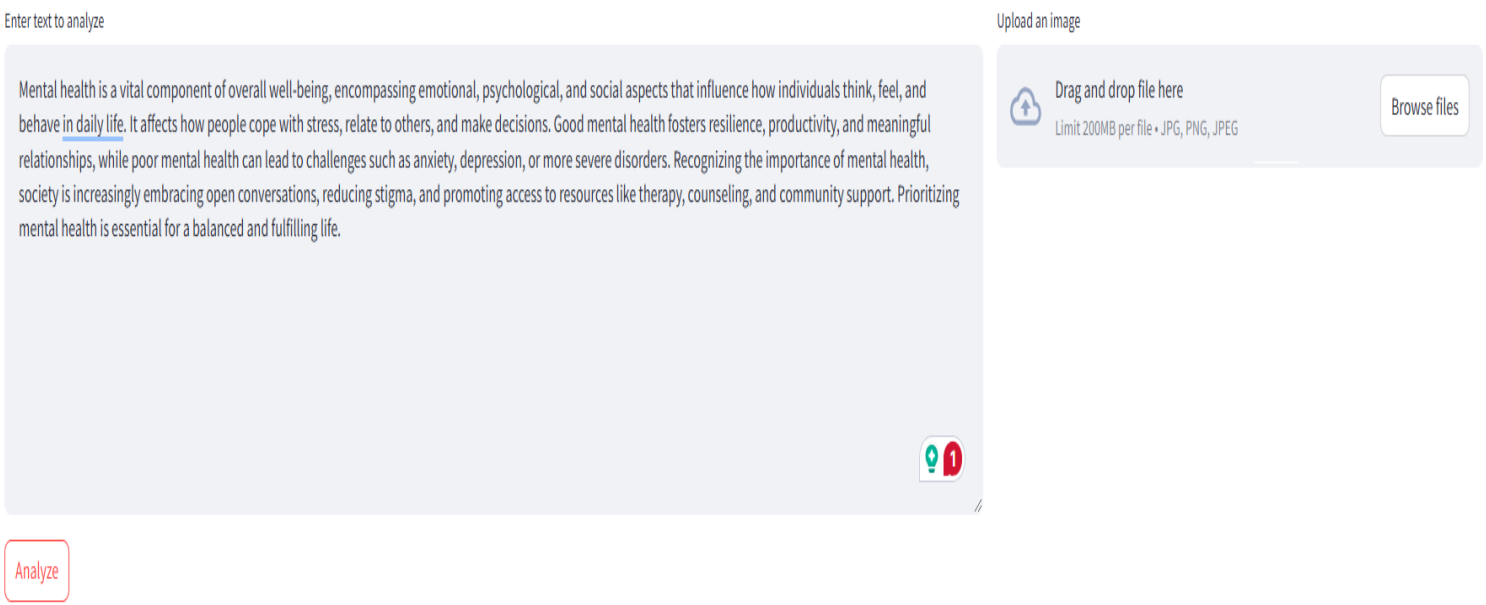
Enter text to analyze
Mental health is a vital component of overall well-being, encompassing emotional, psychological, and social aspects that influence how individuals think, feel, and behave in daily life. It affects how people cope with stress, relate to others, and make decisions. Good mental health fosters resilience, productivity, and meaningful relationships, while poor mental health can lead to challenges such as anxiety, depression, or more severe disorders. Recognizing the importance of mental health, society is increasingly embracing open conversations, reducing stigma, and promoting access to resources like therapy, counseling, and community support. Prioritizing mental health is essential for a balanced and fulfilling life.
Upload an image
Drag and drop file here
Limit 200MB per file • JPG, PNG, JPEG
Browse files
Analyze
Analysis Results
Plagiarism Score
5.93%
AI Generated Label: Real
High probability of AI-generated content.
AI Generated Confidence
99.83%
Human Written Confidence
0.17%

Figure:01

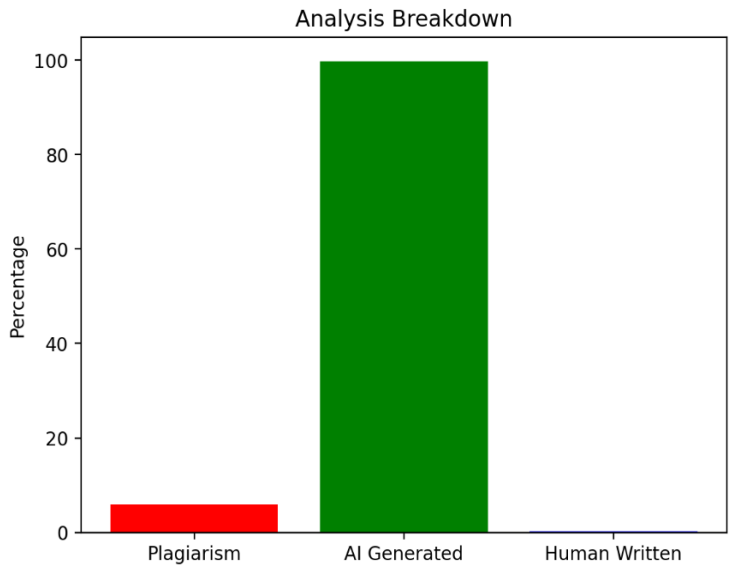

Figure: 02

Analysis History

Analysis 1:

Input Text: AI has changed many fields, including language processing. Tools like GPT-2 have made it easy to cre...

Plagiarism Score: 16.90%

AI Generated: Real (62.49%)

Human Written: 37.51%

Analysis 2:

Input Text: AI has changed many fields, including language processing. Tools like GPT-2 have made it easy to cre...

Plagiarism Score: 16.90%

AI Generated: Real (62.49%)

Human Written: 37.51%

Analysis 3:

Input Text: hello , i am raihan from dhaka . bangladesh . i'm a student in bangladesh university of business and...

Plagiarism Score: 0.00%

AI Generated: Real (99.98%)

Figure :02

CH5 - Conclusion

5.1 Summary: Real Cat successfully integrates state-of-the-art technologies to deliver a robust tool for detecting plagiarism and AI-generated content. The combination of OCR, NLP, and machine learning provides accurate, reliable results.

5.2 Future Work:

5.2.1 Incorporating support for multiple languages in text analysis.

5.2.2 Enhancing OCR accuracy for low-quality images.

5.2.3 Developing real-time analysis for large datasets.

5.2.4 Introducing user-customizable comparison datasets for plagiarism checks.

5.3 Implications: Real Cat holds potential for widespread adoption in academia, publishing, and corporate sectors, ensuring content originality and authenticity in a rapidly evolving digital landscape.

References

1. Vaswani, A., et al. (2017). "Attention Is All You Need." Advances in Neural Information Processing Systems.
2. Devlin, J., et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805.
3. Tesseract OCR Documentation. Retrieved from https://github.com/tesseract-ocr
4. OpenAI API Documentation. Retrieved from https://platform.openai.com
5. Scikit-learn: Machine Learning in Python. Retrieved from https://scikit-learn.org
6. MySQL Database Documentation. Retrieved from https://dev.mysql.com/doc/