Catch The AI

Romani Nasrat, Mohannad Ayman, Mohammed Abdullah, Abdullah Mohammed, Zeyad Elsayed, Ahmed Mohammed, Sara Reda, Rawan Abdel-Aziz, Reham Mostafa

Prof. Eman Abdel-Latef

Faculty of Computers & Artificial Intelligence

Benha University

ABSTRACT

With the rapid advancement of AI technologies, the generation of synthetic media—such as audio, text, and images—has become increasingly sophisticated, blurring the lines between what is real and what is artificially created. In response to the growing challenges posed by AI-generated content, our project focuses on developing a unified approach for detecting AI-generated media across different modalities.

The objective of this study is to empower users with tools that can reliably differentiate between authentic and AI-generated content. We explore three distinct models tailored to analyze text, audio, and images. For text detection, we adopt an ensemble model leveraging RoBERTa and DeBERTa architectures, enhancing feature extraction and prediction accuracy. In the realm of audio, we employ fine-tuning techniques on the Wav2Vec2 model to capture nuanced patterns indicative of AI generation. Lastly, our image detection model integrates state-of-the-art convolutional neural networks (CNNs), including EfficientNet, to effectively discern synthetic images from genuine ones.

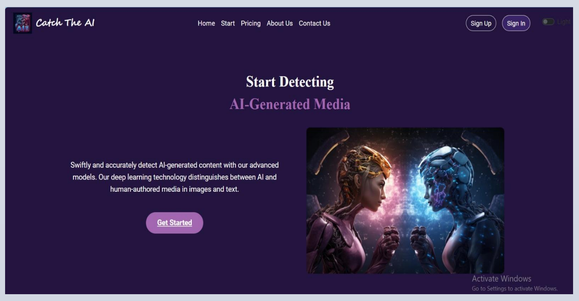
Throughout our research, we faced challenges such as dataset quality, overfitting, and model selection, which we systematically addressed through rigorous experimentation and refinement. Performance evaluations demonstrate promising results across all three models, showcasing high accuracy and robustness in distinguishing between real and AI-generated media.

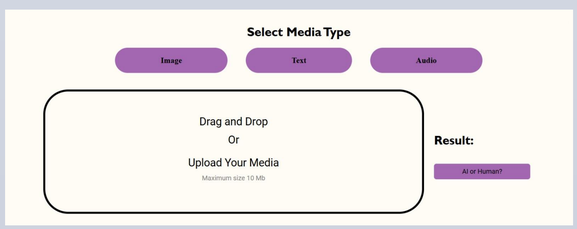
The implications of our project extend to various applications, including media verification, content moderation, and safeguarding against misinformation. Looking ahead, future work entails further refinement of models, exploration into video-based detection, and adaptation for real-time deployment.

In conclusion, our study presents a comprehensive framework for combating the proliferation of AI-generated media, offering a foundational step towards enhancing trust and reliability in digital content consumption.

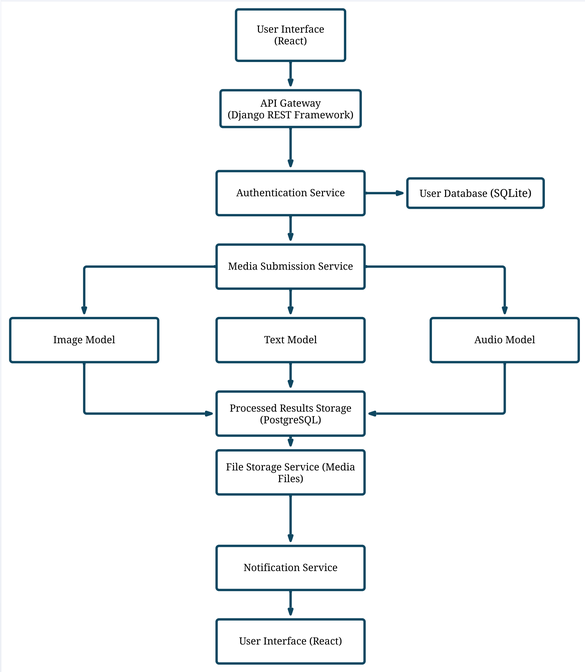
OBJECTIVE

The objective of this project is to develop and implement a unified framework for detecting AI-generated media—specifically audio, text, and images—aimed at empowering users with robust tools to distinguish between authentic and artificially generated content. By leveraging advanced machine learning models and techniques tailored to each modality, our goal is to achieve high accuracy and reliability in identifying instances of synthetic media, thereby contributing to enhanced trust, transparency, and security in digital content consumption and dissemination. Through systematic experimentation and refinement, we aim to provide a scalable solution capable of addressing the evolving challenges posed by AI-generated media across various applications, including media verification, content moderation, and combating misinformation.

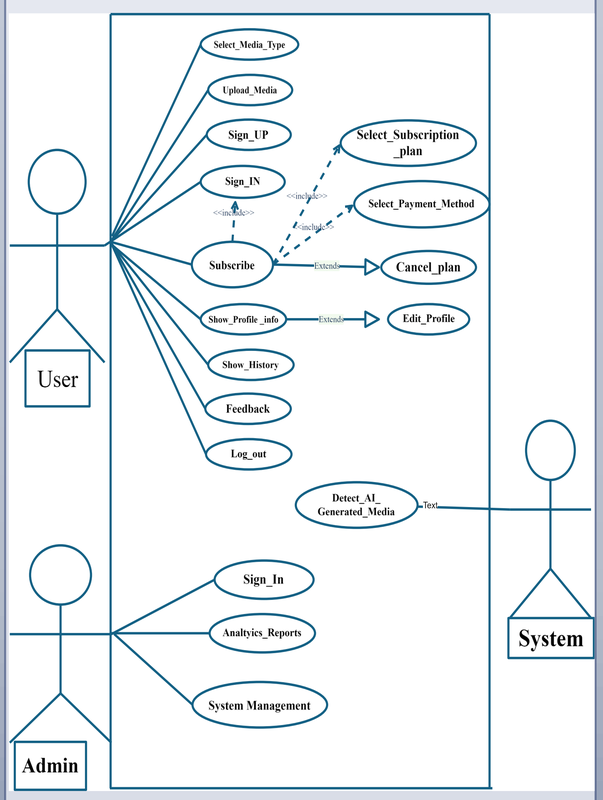




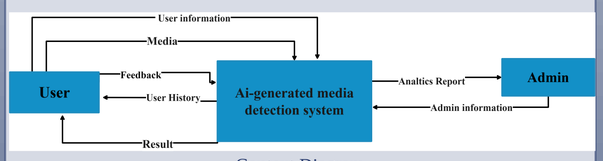
METHODS



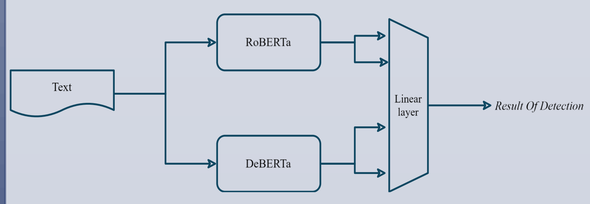
System Architecture Diagram



Use Case Diagram



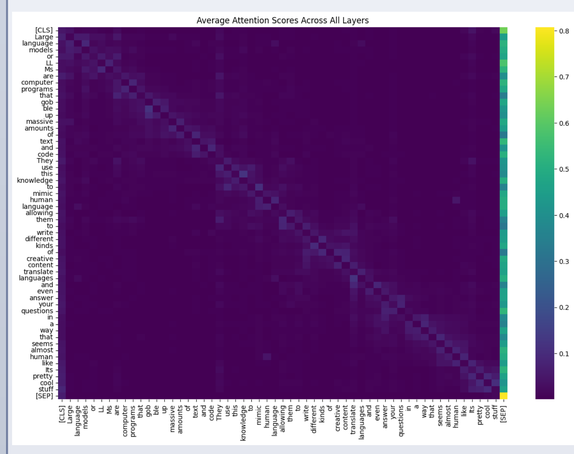
Context Diagram



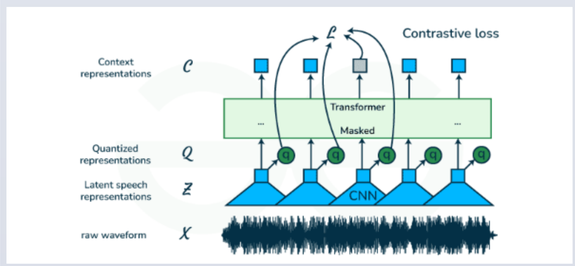
Text Model Architecture

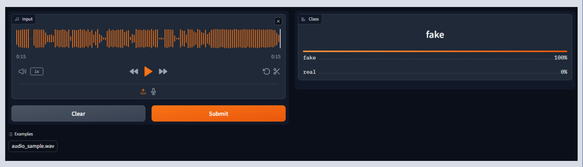
EXPECTATIONS RESULTS

For text-based AI-generated media detection, the system should achieve a high accuracy and precision rate, distinguishing AI-generated text from human-written content with at least 95% precision. This involves analyzing syntactic and semantic patterns, stylistic inconsistencies, and metadata anomalies. Real-time detection is crucial, allowing users to receive immediate feedback on the authenticity of the text content they encounter. The system must also adapt to evolving text generation techniques, ensuring robustness against increasingly sophisticated AI models.

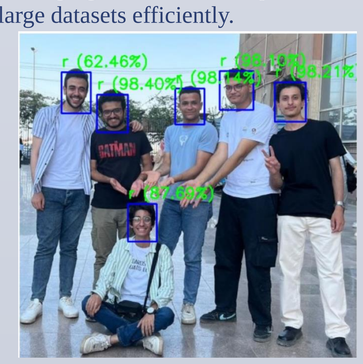


In the realm of audio detection, the system should accurately differentiate between AI-generated and human-generated audio with a similar precision rate. Key indicators might include unnatural speech patterns, inconsistencies in voice modulation, and anomalies in the audio waveform. Real-time processing is particularly important for applications such as live broadcasting or podcasting, where immediate detection of synthetic audio can prevent the spread of misinformation. Scalability is essential, enabling the system to handle large volumes of audio data without performance degradation.





For image-based AI-generated media detection, the system should excel in identifying AI-generated images, including deepfakes and other synthetic visuals, with high precision. Detection techniques should focus on identifying subtle artifacts, irregularities in textures, and inconsistencies in lighting or shadows that are indicative of AI generation. Real-time or near-real-time detection capabilities are vital for applications such as social media monitoring and content moderation. The system must be robust against advanced image generation techniques and scalable to process large datasets efficiently.



MATERIALS

Hugging Face, PyTorch, Django, Bootstrap, Python, React, librosa, Weights & Biases, Visual Studio Code, kaggle, OpenCV

CONCLUSIONS

In this project, we successfully developed an AI-generated media detection web application using React for the frontend and Django for the backend. By leveraging PyTorch and Hugging Face, we fine-tuned models like DeBERTa, RoBERTa, Wav2Vec2, and ViT on our self-collected data. These models were integrated into the application via the Hugging Face serverless API, enabling real-time analysis of user-provided audio, text, and image data.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 2017, pp. 6000-6010. Available: https://arxiv.org/abs/1706.03762.

[2] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS), Vancouver, Canada, 2020. Available: https://arxiv.org/abs/2006.11477.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, 2020. Available: https://arxiv.org/abs/2010.11929.

[4] "Fine-Tuning Transformers with Hugging Face," Hugging Face Documentation, Accessed: Jun. 20, 2024. [Online]. Available: https://huggingface.co/docs/transformers/training.

ACKNOWLEDGEMENTS

We extend our sincere gratitude to our supervisor, Prof. Eman Abdel-Latef, for her invaluable guidance, support, and encouragement throughout this project. Her expertise and feedback were crucial to our success.

We also thank our peers and colleagues for their insights and support, as well as the faculty and staff of BFCAI for providing essential resources. Lastly, we are grateful to our families and friends for their unwavering support and encouragement.