TrueDetect: AI-Powered Media Authenticity Verification

Shuvishka M Sajjan (12)
Avasarala Hiranmayi (13)
M Sanjana (41)
Bonam Sai Sreya (61)
Joanna Grace Fernandez (65)

Problem Statement

As deepfakes and AI-generated audios, images, and text become increasingly indistinguishable from authentic content, the threat of misinformation, fraud, and manipulation grows exponentially. The lack of a reliable mechanism to verify digital media integrity leaves individuals, businesses, and governments vulnerable to deception.

This project aims to address this urgent challenge by developing a cutting-edge tool that accurately detects and verifies the authenticity of digital media, empowering users to trust the information they consume and share.

Literature Review

1

Sha Li, Xinyi Zhang, Ziang Xiao, Curran Kelleher, Diyi Yang (28 Mar 2024) [1]

- Traditional methods like logistics regression, random forest, Multinomial NB, SGDClassifier, SVM, VotingClassifier
- **Semantic Variance Detection**: Deeper word representations like word2vec were effective in capturing subtle semantic differences between machine-generated and human-generated texts

2

Neha Sandotra, Bhavna Arora(14 Dec 2023) [2]

- Uses DL(CNN, LSTM,RNN), ML(SVM, KNN, random forest and decision tree) and statistical learning(3D morphable model)
- Encoders-decoders and generative adversarial network(GANs)
- **Spacial** based features, **temporal** based features, **frequency** based features

3

Haoxin Ma, Jiangyan Yi, Chenglong Wang, Xinrui Yan , Jianhua Tao, Tao Wang, Shiming Wang, Ruibo Fu (Oct 2024) [3]

- **CFAD**: A Chinese dataset for fake audio detection under complex conditions [4]
- More **diverse** real and fake audio types
- Recognizing **unseen** fake algorithms

Use Cases

Text

- Identifying AI-generated fake product reviews on e-commerce platforms.
- Detecting AI-generated comments and bot activity on social media.
- Verifying the authenticity of AI-written articles in journalism.
- Flagging AI-generated job applications and resumes.
- Identifying AI-generated legal documents to prevent fraudulent claims.
- Detecting AI-generated customer service interactions or responses.
- Flagging AI-generated investment advice or financial scams.
- Identifying AI-generated impersonation attempts in email communications

Audio

- Verifying the authenticity of audio used as legal evidence.
- Controlling the spread of misinformation through AI-generated audio clips.
- Preventing defamation or reputational harm to public figures via fake audio.
- Detecting AI-generated voice phishing (vishing) attempts.
- Identifying AI-generated audio in customer service interactions or fraud attempts.
- Flagging AI-generated impersonation in phone calls.
- Verifying the authenticity of AI-generated podcast or media interviews.
- Preventing AI-generated voice in deepfake audio scams targeting businesses or individuals.

Research Objectives

- **Develop Detection Algorithms**: Design advanced algorithms to identify AI-generated text, images, and audio with high accuracy.
- **Assess Accuracy and Reliability:** Evaluate the detection tool’s effectiveness across diverse synthetic media formats and platforms.
- **Enhance Real-Time Detection Capabilities:** Enable real-time identification and flagging of deepfake content to prevent misinformation spread.
- **Optimize for Scalability:** Ensure the tool can handle large volumes of data efficiently for widespread use.
- **Promote Ethical Standards:** Encourage ethical AI practices by embedding measures to prevent harmful uses of synthetic content.

Research Methodology

Text Detection

Objective:
Build a model to distinguish AI-generated text from human-written content, such as detecting whether a student or an LLM (Large Language Model) wrote an essay.

Dataset:
Used the LLM-Detect AI Generated Text Dataset to train the model for accurate classification.

Model Training:
- Logistic Regression: Trained to classify text using TF-IDF features.
- Random Forest: Leveraged ensemble learning on TF-IDF features.
- SVM (Support Vector Machine): Achieved the highest F1 score among all models, making it the final choice for deployment.
- BERT (Transformer Model): Fine-tuned BERT to explore deep learning performance but found SVM more suitable based on evaluation metrics.

Tech Stack:
- Scikit-learn: For model training and evaluation.
- Transformers Library: To experiment with pre-trained models like BERT.
- Streamlit: For building an interactive web application to showcase the prediction results.

Research Methodology

Audio Detection

Objective:
Build a model to detect whether an audio clip is AI-generated (deepfake) or real human speech.

Dataset:
Utilized datasets like TIMIT-TTS for synthetic audio and LibriSpeech for real audio samples.

Model Training:

- XGBoost: gradient-boosting algorithms have outperformed in overall metric scores.

Tech Stack:

- Librosa and IPython for extracting audio features.
- Scikit-Learn for model development.
- Streamlit for building interactive applications to display audio detection results.

Interface & Working

The next few slides showcase screenshots of the working prototype, demonstrating the key features and functionality of the product in action.

Sign in

streamlit_text

localhost:8501

Deploy

AI vs Human Text and Audio Classifier

Text Classification | Audio Classification

Classify Text as AI-Generated or Human-Written

Text Input

Classify Text

Sign in

streamlit_text

localhost:8501

Deploy

AI vs Human Text and Audio Classifier

Text Classification | Audio Classification

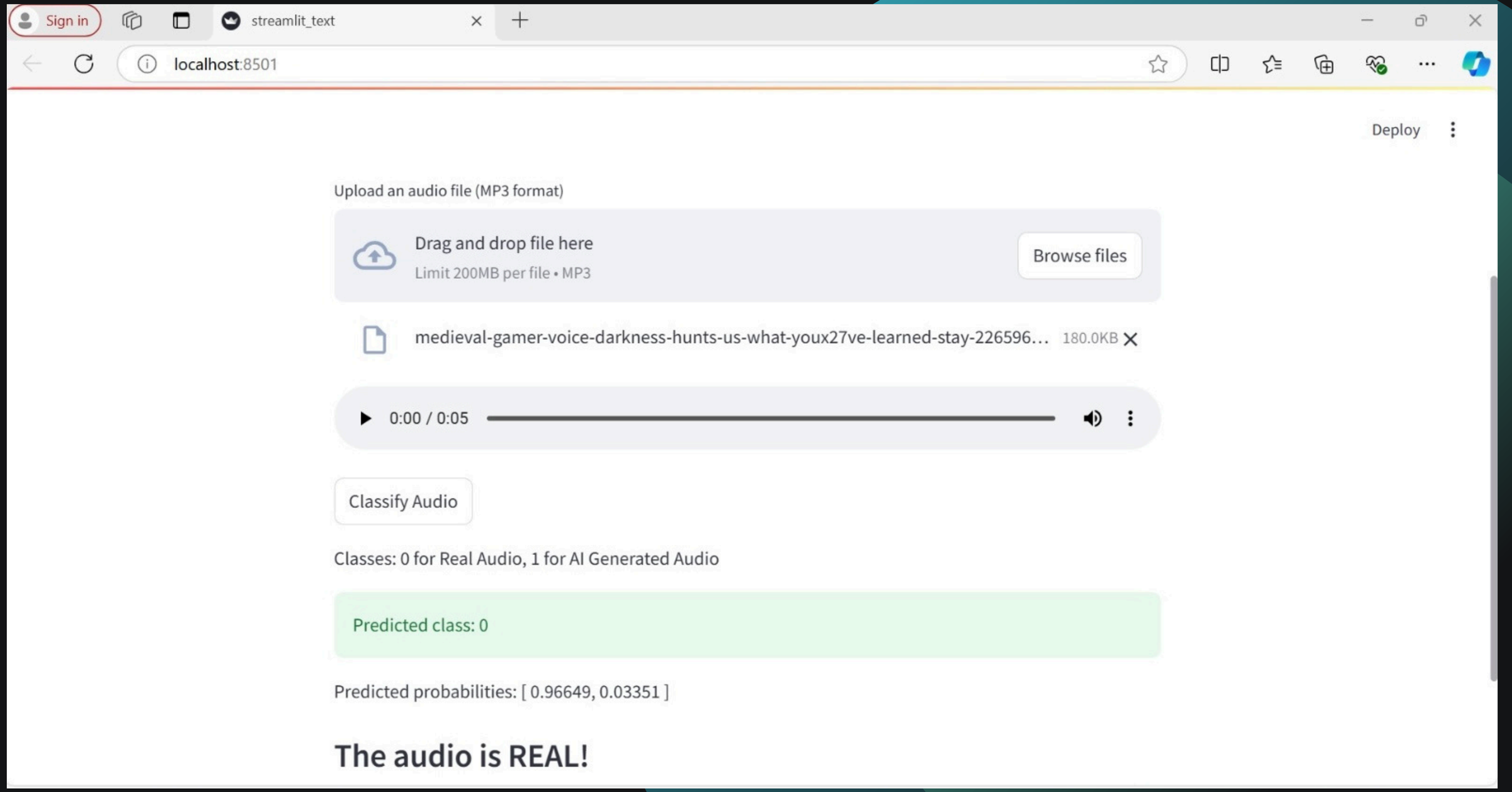
Classify Text as AI-Generated or Human-Written

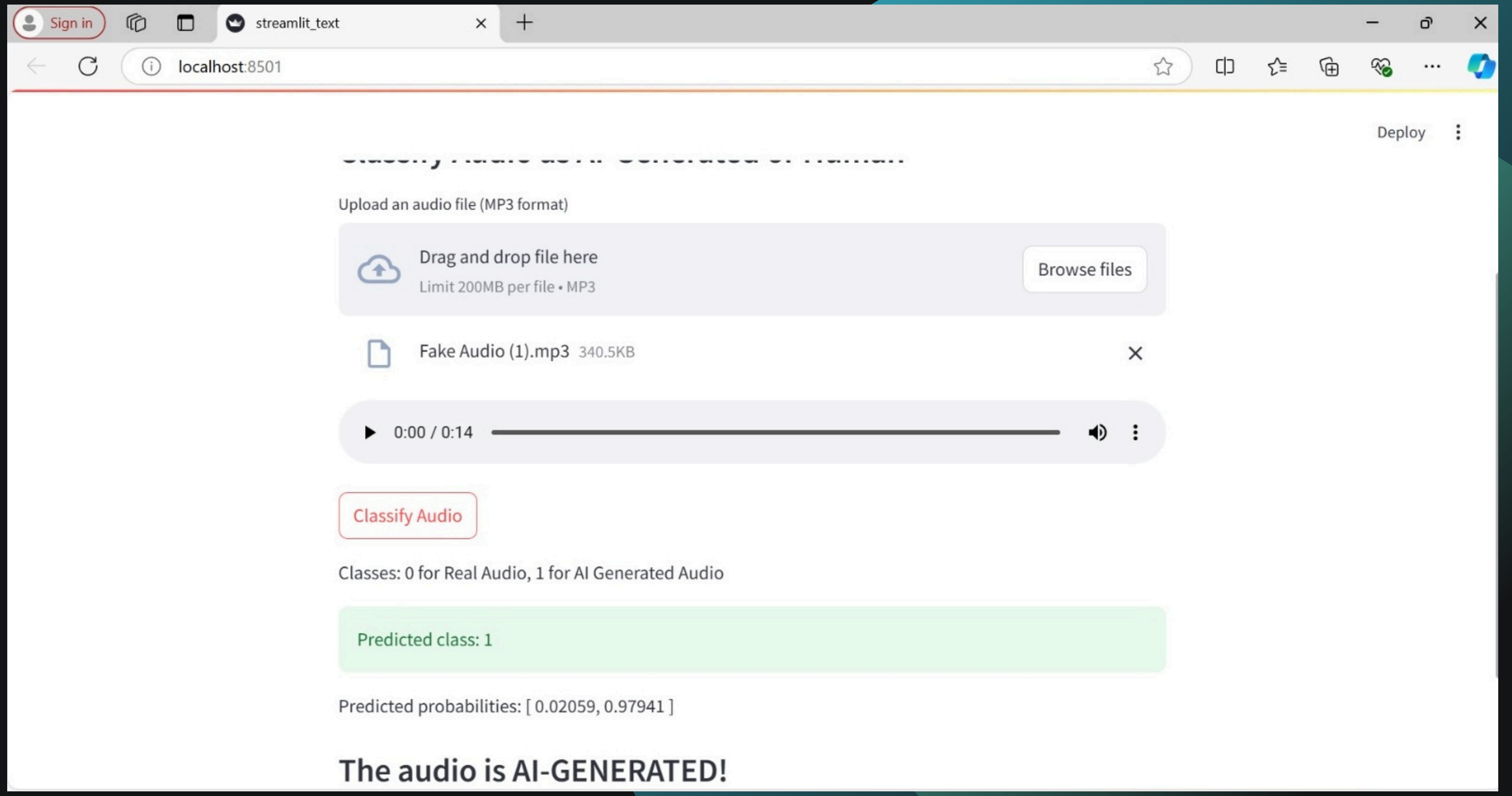
Text Input

Good morning, everyone. I'll be introducing you to face recognition, which is a technology used to identify or verify individuals by analyzing facial features such as the eyes, nose, and jawline. The most common techniques include Eigenfaces, which uses PCA to reduce the complexity of facial data; Active Appearance Models, which combine shape and texture to track facial variations; and 3D Shape Models that use depth to recognize faces from different angles.

Classify Text

The text is AI-Generated!


Sign in
streamlit_text
localhost:8501
Deploy
Upload an audio file (MP3 format)
Drag and drop file here
Limit 200MB per file • MP3
Browse files
medieval-gamer-voice-darkness-hunts-us-what-youx27ve-learned-stay-226596…
180.0KB
0:00 / 0:05
Classify Audio
Classes: 0 for Real Audio, 1 for AI Generated Audio
Predicted class: 0
Predicted probabilities: [0.96649, 0.03351]
The audio is REAL!


Sign in
streamlit_text
localhost:8501
Deploy
Upload an audio file (MP3 format)
Drag and drop file here
Limit 200MB per file • MP3
Browse files
Fake Audio (1).mp3 340.5KB
0:00 / 0:14
Classify Audio
Classes: 0 for Real Audio, 1 for AI Generated Audio
Predicted class: 1
Predicted probabilities: [0.02059, 0.97941]
The audio is AI-GENERATED!

Conclusion

Our aim is to pioneer robust solutions for detecting AI-generated content across text and audio deepfakes. By leveraging cutting-edge techniques and fostering interdisciplinary collaboration, we are dedicated to tackling the increasing challenges posed by AI in content creation.

Through this initiative, we seek to contribute to a more secure digital ecosystem, enhancing authenticity and trust in media, and empowering users to confidently navigate the digital landscape.

References

1. Xie, Y., Rawal, A., Cen, Y., Zhao, D., Narang, S.K. and Sushmita, S., 2024. MUGC: Machine Generated versus User Generated Content Detection. arXiv preprint arXiv:2403.19725.
2. Sandotra, N. and Arora, B., 2024. A comprehensive evaluation of feature-based AI techniques for deepfake detection. Neural Computing and Applications, 36(8), pp.3859-3887.
3. Ma, H., Yi, J., Wang, C., Yan, X., Tao, J., Wang, T., Wang, S. and Fu, R., 2024. CFAD: A Chinese dataset for fake audio detection. Speech Communication, 164, p.103122.
4. https://github.com/ADDchallenge/CFAD
5. https://www.kaggle.com/datasets/sunilthite/llm-detect-ai-generated-text-dataset/data
6. https://www.kaggle.com/datasets/birdy654/cifake-real-and-ai-generated-synthetic-images
7. https://www.kaggle.com/datasets/birdy654/deep-voice-deepfake-voice-recognition
8. Frank, J. and Schönherr, L., 2021. Wavefake: A data set to facilitate audio deepfake detection. arXiv preprint arXiv:2111.02813.

Thank You!