Synthetic Misinformation Detection: A Multimodal System for Identifying Fake News and Deepfakes

Introduction & Problem Statement

In the digital era, the spread of misinformation is no longer confined to text—it now includes realistic synthetic videos and AI-generated content that can mislead and manipulate public opinion at scale. Deepfakes (AI-generated fake videos) and fake news articles have become powerful tools for misinformation campaigns, posing severe threats to politics, public health, and societal trust.

This project aims to develop a **multimodal AI system** that detects both:

1. **Textual misinformation (Fake News)**
2. **Visual misinformation (Deepfakes)**

I will combine traditional machine learning and deep learning techniques with recent advances in **Large Language Models (LLMs)** and **CNN-based deepfake detectors**. The final system will be able to **analyze both text and video** and classify whether content is real or fake, providing **explanations** using LLMs and **confidence scores** based on image/video model output.

Data Sources

Text (Fake News):

- **LIAR dataset**: 12.8k short political statements with 6-level truth labels from PolitiFact.
- **FakeNewsNet**: Includes full news articles, tweet context, and credibility labels.
- **FEVER** (optional): Fact-checking dataset with claims and evidence for claim verification.

Video/Image (Deepfakes):

- **FaceForensics++**: 1,000+ real/fake videos generated using multiple deepfake methods.
- **Deepfake Detection Challenge (DFDC)**: Large-scale dataset with over 100,000 videos labeled real/fake.

Methods, Technologies & Techniques

Fake News Detection (Text):

- **Preprocessing**: Tokenization, text cleaning, and formatting.
- **Modeling Approaches**:
    - Classical ML (TF-IDF + Logistic Regression/Random Forest)
    - Deep Learning (Fine-tuned BERT or RoBERTa classifier)
    - **LLM-based claim verification** using Claude:
        - Prompted to evaluate truthfulness and return structured explanations and confidence.

Deepfake Detection (Video/Image):

- **Preprocessing**: Face extraction and frame sampling from videos.
- **Modeling Approaches**:
    - Pretrained CNNs like **XceptionNet** and **MesoNet** fine-tuned on FaceForensics++
    - Optional: frame-level ensembling or temporal attention mechanisms
    - Evaluation using F1, accuracy, and confusion matrix

Multimodal Fusion (Stretch Goal):

- Combine text and video model outputs into a joint confidence score or use **CLIP-style embeddings** for cross-modal similarity and classification.

Deliverables

1. **Deepfake Detection Module (Main focus of project)**:
    - CNN-based model that classifies input video/image as real or fake.
    - Frame-level and video-level predictions with confidence visualization.

2. **Fake News Detection Module (Will have an MVP for this ready)**:

   - Trained classifier (BERT) that outputs real/fake label for articles or headlines.
   - LLM-based verifier (e.g., GPT-3.5) that provides claim reasoning and confidence.
   - Evaluation metrics: Accuracy, F1, ROC curve.

3. **Documentation & Final Report**:
   - Full implementation on GitHub, complete with Jupyter notebooks or app demo.
   - A written report covering background, data, methodology, and results.