Project 2: A Multimodal harmful meme detection System

Project Description
The "Multimodal Harmful Meme Detection System" aims to develop a classifier that can effectively detect harmful content in memes by analyzing both image and text elements. This will involve using multimodal fusion techniques to capture the complex interplay between text and images, improving detection accuracy.

Reference Paper: https://aclanthology.org/2022.nlp4pi-1.20.pdf

Dataset
The **Hateful Memes Challenge (HMC)** dataset would be ideal for this project, as it provides a variety of memes with annotations for hatefulness based on a combination of text and image contexts. This dataset is already widely used for research on harmful meme detection and offers well-curated examples to train and validate a multimodal model.
Dataset link: https://paperswithcode.com/paper/the-hateful-memes-challenge-detecting-hate

Models

1. **Text Encoder**: You can use language model like **BERT** or **DistilBERT** for processing textual content. Fine-tuning on harmful language detection will enhance the system's capability to interpret hateful language in meme text.
2. **Image Encoder**: You can implement a **Vision Transformer (ViT)** or **ResNet**, pretrained on large datasets, for analyzing visual elements. Fine-tuning on meme data will optimize the model for the specific characteristics of meme imagery.
3. **Multimodal Fusion Model**: You can use a fusion method similar to Hate-CLIPper's **Feature Interaction Matrix (FIM)** to combine text and image features effectively. This matrix models the relationships between image and text features, facilitating meaningful cross-modal interactions that are vital for accurately detecting harmful content.
4. You may also consider exploring **Parameter-Efficient Fine-Tuning (PEFT)** methods. These techniques allow you to fine-tune large models efficiently by updating only a subset of parameters, such as using adapters or **Low-Rank Adaptation (LoRA).** PEFT methods can be especially helpful in reducing computational costs while maintaining performance.

   Important: You are free to choose the models you prefer.

Interface
Build an interactive web interface where users can:

- **Upload or Enter URL**: Submit a meme for analysis.
- **View Prediction Results**: Display the meme alongside predictions of harmfulness with confidence scores for each modality.
- You could use **Streamlit** or **Flask** to create a user-friendly interface.