MULTIMODAL SARCASM DETECTION

(Using Facial Expressions, Audio, and Text Processing)

03 October,2024

Md Nazmus Saquib Khan
(2011537 042)
Md Zian Raian
(2011394042)
Kazi Nafisur Rahman
(2013628642)
Mosammat Mariya
(1931148642)

SYSTEM OVERVIEW

- The system integrates:

  Facial Expression: Analysis of facial cues over 5-10 seconds.

  Audio: Analysis of voice patterns (intonation, pitch).

  Text: NLP-based sarcasm detection.

- Multimodal Fusion: Combines outputs from all three modalities to make a final sarcasm prediction.

- Inputs (Face, Audio, Text) → Modality-specific models → Feature Fusion → Final Sarcasm Detection.

FACIAL EXPRESSION MODEL

Model:

- CNN (e.g., VGGFace) to extract facial features.
- LSTM to analyze the temporal dynamics over 5-10 second videos.
- Features: Micro-expressions, smirks, and subtle facial cues.
- Tools: OpenCV.

- Video frames → CNN → LSTM → Feature Vector for sarcasm detection.

AUDIO PROCESSING MODEL

Model:

- MFCC and spectrogram extraction for capturing vocal features.
- LSTM or Transformer for time-series analysis of the audio signal.
- Features: Sarcastic tone, intonation, stress, and pauses.
- Tools: Librosa (for MFCC), Pretrained models like Wav2Vec.

- Audio signal → MFCC/Wav2Vec → LSTM → Feature Vector.

TEXT PROCESSING MODEL

Model:

- Use BERT, RoBERTa for text-based sarcasm detection.
- Fine-tuned on sarcasm-specific data, capturing contradictions and irony in context.
- Features: Irony, exaggeration, incongruity between sentiment and meaning.
- Tools: BERT (via Hugging Face).

- Text input → BERT → Contextual embeddings → Feature Vector.

MULTIMODAL FUSION

Fusion Method:

- Combine feature vectors from face, audio, and text models(Early/Late Fusion).
- Use attention mechanisms to weigh the importance of each modality dynamically (Multimodal Transformer).
- Output: The fused vector is passed through a final classifier for sarcasm prediction.

CONCLUSION & FUTURE WORK

Multimodal sarcasm detection captures both verbal and non-verbal cues, improving accuracy.

- Future Work:
- Improve real-time detection capabilities.
- Extend the model to more languages and larger datasets.
- Impact: A more intuitive interaction between machines and humans in daily life.

DATASET - MUStARD

- Sarcasm-specific multimodal dataset from TV shows.
- Modalities: Video (facial expressions), Audio (vocal intonation), and Text (dialogue transcripts).
- Size: ~690 utterances with sarcasm labels.