MULTIMODAL SARCASM DETECTION USING FACIAL EXPRESSION, VOICE EXPRESSION AND DIALOGUE

Md Nazmus Saquib Khan 2011537 042
Md Zian Raian 2011394 042
Kazi Nafisur Rahman 2013628 642
Mosammat Mariya 1931148 642

What is Our Project

- Overview: A Multimodal Sarcasm Detection System using text, video, and audio to detect sarcasm in multimedia.
- Motivation: Sarcasm is complex, involving tone, expressions, and context—key for improving AI understanding in real time human- Machine interaction.
- Goal: An ensemble model that processes multiple data types to classify videos as sarcastic or non-sarcastic.

Related Work

- Wu et al. worked on detecting sarcasm by focusing on the differences between what people say and how they express it. Their paper, Modeling Incongruity between Modalities for Multimodal Sarcasm Detection, shows how sarcasm is often a mismatch between positive words and negative expressions or tone. They developed a model that looks at text, voice, and facial expressions together
- Castro et al., in Towards Multimodal Sarcasm Detection (An Obviously Perfect Paper), used a more advanced method. They combined text, voice, and facial expression data with an attention mechanism, which helped their model focus on the most important parts of each. This allowed their system to catch sarcasm more effectively by weighing the information from each source differently
- Pramanick et al., in their paper Multimodal Learning using Optimal Transport for Sarcasm and Humor Detection, took a different approach. They used a method called optimal transport to ensure that features from text, voice, and facial expressions worked together smoothly. This helped their model detect both sarcasm and humor more effectively.

Social Effects

- Positive Impact: Enhances communication by reducing misunderstandings in social media, customer support, and human-robot interactions.
- Potential Risks: Sarcasm is nuanced and culturally specific, so errors could cause misinterpretations and frustration.
- Mitigation: Controlled deployment, feedback-based updates, and providing suggestions rather than absolute judgments to improve adaptability.

Environmental Effects

- Resource Use: Training multimodal deep learning models demands high computational power, increasing energy consumption and carbon footprint.
- Mitigation: We will use pre-trained models (e.g., Resnet, BERT) to reduce training needs, and select energy-efficient cloud solutions. A lot of our work is already done.
- Sustainability Practices: Leveraging cloud services that use renewable energy and carbon-neutral data centers to reduce environmental impact.

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.Resnet) to extract facial features.
- Features: Micro-expressions, smirks, and subtle facial cues.
- Tools: OpenCV.
- Video frames → CNN → LSTM → Feature Vector for sarcasm detection

AUDIO PROCESSING MODEL

- AUDIO PROCESSING MODEL
- Model: Wave2vec
- Features: Sarcastic tone, intonation, stress, and pauses.
- Tools: Librosa (for MFCC).
- Audio signal → MFCC/Wav2Vec → LSTM → Feature Vector

TEXT PROCESSING MODEL

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

MULTIMODAL FUSION

- Fusion Method:
- Combine feature vectors from face, audio, and text models(Early/Late Fusion).
- Ensemble Method:
- Taking votes from different model and output will be based on voting.

High Level Technical Design

- Using different pretrained and custom made model for each modalities
- Trying Combination of different Models
- Training Different Datasets for Different Modalities if Possible
- Ensemble or Feature fusion for Final Model

Our Progress

- Dataset Processing Department
- Dataset for TEXT
- Dataset for Audio
- Dataset for video

Our Progress

- Model Building Department
- TEXT Model Based on BERT
- Audio Model Based On WAVE2VEC
- Video Model Based on Resnet18
- Very Close to 1st ensemble model
- Halfway progress on 1st feature fusion model.