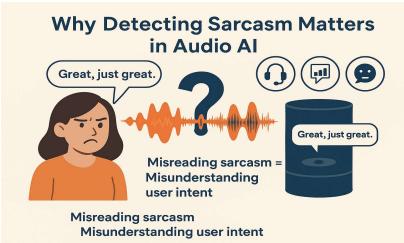
No Words , Just Tone : Audio-Based Sarcasm Detection

Aditya Kumar Sinha* , Nikhil Kumar* ,Orchid Chetia Phukan, Arun Balaji Buduru
{aditya22034, nikhil22322)@iiitd.ac.in IIITD

Abstract

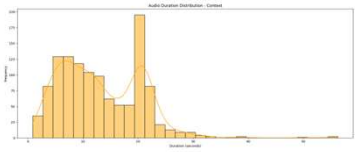
Sarcasm is a nuanced form of expression often misinterpreted by sentiment-based AI systems. While prior research emphasizes multimodal analysis, we explore whether audio alone can effectively detect sarcasm. Using state-of-the-art pretrained models like wav2vec2, HuBERT, Whisper, and others, we extract audio embeddings from the MUStARD++ dataset. Our deep learning models—CNN+FCN and FCN-only—achieve state-of-the-art performance, surpassing multimodal baselines. This challenges the reliance on multimodal inputs and highlights the potential for efficient, unimodal sarcasm detectors.



Dataset

The MUStARD++ dataset is a multimodal benchmark curated for sarcasm detection research. It comprises short clips from popular sitcoms like *Friends* and *The Office*, annotated with binary sarcasm labels. Each sample includes four synchronized modalities: text, audio, video, and conversational context. The dataset is designed to capture the subtle cues of sarcasm across different media formats. For this work, we focus exclusively on the audio modality, using the "audio utterance" (target line) and "audio context" (preceding dialogue) components. These rich and diverse clips offer a realistic foundation for evaluating unimodal sarcasm detection systems in spoken language.

The context audio, with an average duration of 13.18 seconds, exhibits a broader range of lengths, reflecting variable dialog history.



Methodology

Audio Embedding Extraction
Audio clips from the MUStARD++ dataset were passed through various pretrained models such as Wav2Vec2.0, HuBERT, Whisper, MMS, and LanguageBind to extract fixed-length embeddings. Each sample included both context and utterance audio, and embeddings from single or paired models were used.

Model Architectures
Four deep learning models were explored:
1.FCN – Dense layers for context and utterance embeddings, later fused for classification.
2.CNN+FCN – Embeddings reshaped and passed through Conv1D layers to capture local patterns.
3.Dual-Embedding FCN – Independent FCNs process embeddings from two models, then fuse for final prediction.
4.Dual-Embedding CNN+FCN – Combines Conv1D layers and dense layers across two model embeddings for richer feature learning.

Training Strategy
All models used binary cross-entropy loss and Adam optimizer. Early stopping and checkpointing were applied to avoid overfitting. Evaluation was based on accuracy, precision, recall, and F1-score.

Results

Performance on Single-Model Embeddings (CNN vs FCN)

MODEL	CNN		FCN	
FEATURES	ACC	F1	ACC	F1
ImgaeBind + Langbind	70.83	70.78	68.33	67.9
ImageBind + Whisper	71.66	71.28	69.16	68.54
ImageBind + MMS	0.7	69.7	71.66	71.01
ImageBind + XLS-R	66.66	65.9	0.7	69.87
LanguageBind + Whisper	73.33	73.15	72.5	72.48
LanguageBind + MMS	71.37	71.32	70.95	70.53
LanguageBind + XLS-R	73.86	73.85	71.37	70.47
Whisper + MMS	70.12	70.06	68.05	68.04
Whisper + XLS-R	69.29	69.28	68.05	67.94
MMS + XLS-R	69.29	69.18	67.63	67.63

Performance on Concatenated Embeddings (CNN vs FCN)

MODEL	CNN		FCN	
FEATURES	ACC	F1	ACC	F1
UNISPEECH	66.8	66.8	66.8	66.79
wav2vec2	67.63	67.61	67.22	67.21
wavlm	68.46	68.31	67.22	66.77
Whisper	**72.61**	72.48	**73.03**	73
XLSR	67.22	66.19	69.29	69.28
xvector	68.05	67.97	68.88	68.25
MMS	69.29	69.14	68.88	68.83
hubert	68.46	66.86	69.29	68.63
LanguageBind	70.95	70.84	71.78	71.43
ImageBind	68.46	68.34	67.22	67.07

Conclusion

Single-Model Embeddings
- Whisper and LanguageBind yielded the highest accuracy and F1-scores across both CNN and FCN; Whisper slightly outperformed in FCN.
- Whisper's multilingual robustness and real-world noise handling, along with LanguageBind's multimodal semantics (capturing tone and intent), made them especially effective for sarcasm detection.
- CNN and FCN performed similarly, with FCN slightly better for time-independent embeddings and CNN better for capturing local sequential patterns.

Concatenated (Dual-Model) Embeddings
- Combining embeddings significantly improved performance, highlighting the benefit of enriched, complementary feature spaces.
- LanguageBind + XLS-R (CNN) achieved the best overall results, followed by LanguageBind + Whisper—effectively merging semantic and acoustic strengths.
- This dual-embedding approach enabled the model to capture both high-level semantics and low-level prosodic cues, essential for detecting sarcasm.

General Trends and Observations
- CNNs were more effective at modeling temporal relationships in high-resolution and concatenated embeddings.
- FCNs performed competitively for single-model embeddings, especially when temporally pooled or flattened.

References

[1] Baevski et al. wav2vec 2.0: Self-supervised speech representations. arXiv:2006.11477, 2020.
[2] Castro et al. Multimodal sarcasm detection. ACL, pp. 4619–4629, 2019.
[3] Conneau et al. Unsupervised cross-lingual speech representation learning. arXiv:2006.13979, 2020.
[4] Girdhar et al. ImageBind: One embedding space to bind them all. arXiv:2305.05665, 2023.
[5] Hsu et al. HuBERT: Self-supervised speech learning via masked prediction. IEEE/ACM TASLP, 2021.
[6] Pratap et al. Scaling speech tech to 1000+ languages. arXiv:2305.13516, 2023.
[7] Radford et al. Robust speech recognition via weak supervision. arXiv:2212.04356, 2023.
[8] Ray et al. Multimodal corpus for emotion recognition in sarcasm. LREC, pp. 6992–7003, 2022.
[9] Snyder et al. X-vectors: DNN embeddings for speaker recognition. ICASSP, pp. 5329–5333, 2018.
[10] Zhang et al. UniSpeech: Unified speech representation learning. arXiv:2101.07597, 2021.[11] Zhao et al. LanguageBind: Extending video-language pretraining to new modalities. arXiv:2305.16414,