Supplementary Material

Acoustic-Prosodic Evidence in Multimodal Sarcasm Detection: A Controlled and Interpretable Evaluation of PEFM-CMAE on the Complete MUStARD Corpus

Iyanuoluwa Modupe Fatoki; Victoria Bolanle Oyekunle; Emmanuel Adediran, PhD (corresponding author)

Correspondence: adediran.emmanuel@lcu.edu.ng

Purpose. This supplement consolidates the corpus audit, architecture specification, experimental configurations, aggregate results, statistical comparisons, modality-gate summaries, cross-show stress tests, reproducibility information and known artefact limitations that support the accompanying article. It contains no fabricated per-instance predictions, gate traces, final-model error examples or hardware logs; those items were not retained by the completed experimental environment.

S1. Supplementary corpus audit

The complete audio-bearing MUStARD preparation contains 690 matched transcript-waveform pairs and is globally balanced, although programme-level label distributions are highly uneven. The programme imbalance is central to interpreting leave-one-show-out results.

Table S1. Show-level composition of the complete corpus

Source show	Non-sarcastic	Sarcastic	Total	Sarcasm share (%)
The Big Bang Theory	140	140	280	50.0
Friends	204	152	356	42.7
The Golden Girls	1	39	40	97.5
Sarcasmaholics Anonymous	0	14	14	100.0
Total	345	345	690	50.0

Note. The two smallest programmes are almost or completely single-class; their held-out macro-F1 values are stress-test indicators rather than stable model rankings.

Table S2. Cross-show stress-test results and label context

Held-out show	N	Sarcasm share (%)	Text only	Text + audio	PEFM-AE2	PEFM-CMAE
Friends	356	42.7	0.6039	0.6489	0.6727	0.6713
The Big Bang Theory	280	50.0	0.5496	0.6107	0.6393	0.6426
The Golden Girls	40	97.5	0.4937	0.5833	0.4805	0.4667
Sarcasmaholics Anonymous	14	100.0	0.4815	0.3636	0.4167	0.4167

Note. Show identity changes speaker sets, dialogue style, production conditions and label priors simultaneously. This protocol therefore measures joint programme-domain shift, not a clean leave-one-speaker-out effect.

S2. Detailed architecture and interaction equations

The text branch uses cached 768-dimensional RoBERTa-base features; the acoustic branch uses cached mean-pooled 768-dimensional HuBERT-base features; the emotion branch uses a four-element posterior vector over neutral, happy, angry and sad. The text and audio vectors are projected from 768 to 256 dimensions, while the emotion posterior is projected from 4 to 256 dimensions.

The explicit text-audio incongruity descriptor is formed as $d = [|t - a| ; t \odot a]$, where $|t - a|$ encodes projected disagreement and $t \odot a$ encodes element-wise interaction. The 512-dimensional descriptor is mapped through an MLP of dimensions $512 \rightarrow 256 \rightarrow 256$ to obtain the incongruity vector i.

The modality gate receives $[t ; a ; e ; i]$ (1024 dimensions), applies a linear transformation from 1024 to 3, and uses softmax to produce text, audio and emotion weights. The weighted fusion $f = \alpha t\ t + \alpha a\ a + \alpha e\ e$ is concatenated with i, giving $z = [f ; i]$ (512 dimensions). The classifier maps z through a 512-unit hidden layer to two binary-class logits.

Table S3. Dimensional specification of PEFM-CMAE

Component	Input dimension	Transformation	Output dimension
Text projection	768	Linear + activation	256
Audio projection	768	Linear + activation	256
Emotion projection	4	Linear + activation	256
Incongruity descriptor	512	$[\|t - a\| ; t \odot a]$	512
Incongruity MLP	512	$512 \rightarrow 256 \rightarrow 256$	256
Modality gate	1024	Linear $1024 \rightarrow 3$ + softmax	3 weights
Fused vector	3×256	Weighted sum	256
Classifier input	256 + 256	$[f ; i]$	512
Classifier	512	Hidden $512 \rightarrow 2$	2 logits

S3. Experimental configurations and hyperparameters

Table S4. Controlled model configurations

Configuration	Active information	Comparative purpose
Text-only RoBERTa	Text + context	Lexical baseline
Audio-only HuBERT	Waveform	Acoustic-prosodic baseline

Text + audio	Text and waveform	Strong direct bimodal baseline
Text + audio + acoustic emotion	Three-stream concatenation	Tests the affective stream without learned gating
PEFM-AE2	Three streams + softmax gate	Tests adaptive modality weighting
PEFM-CMAE	Incongruity block + softmax gate	Tests explicit incongruity and adaptive gating against direct fusion

Table S5. Core experimental settings

Setting	Value
Text encoder	RoBERTa-base; frozen cached 768-dimensional features
Acoustic encoder	facebook/hubert-base-ls960; frozen; mean-pooled cached 768-dimensional features
Emotion stream	Frozen four-class HuBERT-based posterior over neutral, happy, angry and sad
Projection dimensions	Text/audio 768 → 256; emotion 4 → 256
Classifier	[f ; i] (512) → hidden layer 512 → 2 logits; dropout applied
Optimiser	AdamW; weight decay 0.01
Fusion/head learning rate	1×10^{-3}
Batch size	4
Maximum epochs	30
Early stopping	Validation macro-F1; patience 8; heuristic fixed choice rather than search result
Loss	Cross-entropy; fold class weights were 1.0 and 1.0 because stratification gave balanced training partitions
Text limit	192 subword tokens; longer sequences right-truncated
Audio limit	8 seconds at 16 kHz; longer waveforms right-truncated
Primary protocol	5 folds × 3 seeds = 15 paired partitions
Stress test	Leave one source show out

S4. Complete-corpus results

Table S6. Stratified performance across fifteen seed-fold partitions

Configuration	Accuracy	Macro-precision	Macro-recall	Macro-F1
Text-only RoBERTa	0.6918 ± 0.0344	0.6939 ± 0.0344	0.6918 ± 0.0344	0.6909 ± 0.0346
Audio-only HuBERT	0.7295 ± 0.0406	0.7307 ± 0.0409	0.7295 ± 0.0406	0.7291 ± 0.0406
Text + audio	0.7483 ± 0.0400	0.7501 ± 0.0407	0.7483 ± 0.0400	0.7479 ± 0.0400

Text + audio + acoustic emotion	0.7440 ± 0.0328	0.7453 ± 0.0330	0.7440 ± 0.0328	0.7436 ± 0.0329
PEFM-AE2	0.7329 ± 0.0343	0.7340 ± 0.0343	0.7329 ± 0.0343	0.7325 ± 0.0344
PEFM-CMAE	0.7512 ± 0.0332	0.7543 ± 0.0332	0.7512 ± 0.0332	0.7504 ± 0.0334
PEFM-CMAE minus text + audio	+0.0029	+0.0042	+0.0029	+0.0025

Table S7. Paired macro-F1 comparisons on identical partitions

Comparison	Mean A	Mean B	Δ	95% CI for Δ	t	p	Cohen d
PEFM-CMAE vs text-only RoBERTa	0.7504	0.6909	0.0595	0.0432 to 0.0758	7.8531	0.000002	2.03
PEFM-CMAE vs PEFM-AE2	0.7504	0.7325	0.0179	0.0086 to 0.0272	4.1097	0.001062	1.06
PEFM-CMAE vs text + audio	0.7504	0.7479	0.0025	−0.0106 to 0.0156	0.4088	0.688882	0.11
PEFM-CMAE vs three-stream concatenation	0.7504	0.7436	0.0068	−0.0045 to 0.0181	1.2881	0.218594	0.33
Text + audio vs text-only RoBERTa	0.7479	0.6909	0.0570	0.0374 to 0.0766	6.2500	0.000021	1.61

Note. The PEFM-CMAE advantage over direct text-plus-audio fusion was not statistically significant. Effect sizes are paired fold-level standardised differences.

Table S8. Learned modality gate weights

Configuration	Text gate	Audio gate	Emotion gate
PEFM-AE2	0.6024 ± 0.1893	0.3546 ± 0.1750	0.0430 ± 0.0715
PEFM-CMAE	0.7312 ± 0.1875	0.2357 ± 0.1594	0.0331 ± 0.0494

Note. Gate weights are descriptive allocations, not causal explanations. Per-instance gates were not retained, preventing valid retrospective subgroup analysis.

S5. Equivalence sensitivity analysis

For the PEFM-CMAE versus text-plus-audio contrast, the paired mean difference was 0.0025 macro-F1 and the 95% confidence interval was −0.0106 to 0.0156. A post-hoc two-one-sided-tests sensitivity analysis used ±0.010 macro-F1 as a smallest effect of practical interest because an absolute gain below one macro-F1 point was considered too small to justify the added fusion complexity and was smaller than the observed between-partition variability. TOST p = 0.120 did not establish equivalence. Because the bound was not preregistered, this analysis is descriptive rather than confirmatory.

S6. Contextual comparison with prior MUStARD reports

Table S9. Indicative published results and protocol caveats

Study/model	Modalities	Protocol	Reported metric	Value
Castro et al., SVM	Text	Speaker-independent	F1	0.609
Castro et al., SVM	Text + audio	Speaker-independent	F1	0.631
Castro et al., SVM	Text + audio + video	Speaker-independent	F1	0.643
Chauhan et al., multitask	Text + audio + video	Speaker-independent	Accuracy	0.7694
Gao et al., attention-based fusion	Text + audio	Speaker-independent	F1	0.810
This study, direct fusion	Text + audio	5-fold × 3 seeds	Macro-F1	0.7479
This study, PEFM-CMAE	Text + audio + emotion	5-fold × 3 seeds	Macro-F1	0.7504
Raghuvanshi et al.	Text + audio + video	MUStARD++; study-specific	Macro-F1	0.7496

Note. These values are contextual only. Speaker-independent, speaker-dependent and repeated stratified cross-validation protocols estimate different targets, while F1 variants, class distributions, modalities and augmentation also differ. The rows must not be interpreted as a direct ranking.

S7. Available exploratory error evidence

Table S10. Error categories from one earlier-phase fold

Error category	Count	Share (%)	Interpretation
Context dependency	4	36.4	Evidence beyond the fixed context window
Lexical-cue absence	3	27.3	Little overt textual marking
Prosodic ambiguity	3	27.3	Subtle or deadpan delivery unresolved
Speaker drift	1	9.1	Atypical speaker pattern
Total	11	100.0	Exploratory single-fold review

Note. This review concerns one fold from the earlier 200-utterance phase and is not the final PEFM-CMAE error distribution. Final-model pooled predictions and contexts were not retained, so final-model examples and confusion matrices are not reconstructed here.

S8. Reproducibility statement

- The primary analysis uses the complete 690-utterance audio-bearing MUStARD preparation reconstructed from the official annotations and raw audiovisual clips. Raw copyrighted media are not redistributed.

• Text features were generated with RoBERTa-base. Acoustic features were generated with facebook/hubert-base-ls960 and aggregated by arithmetic mean pooling. The speech-emotion stream is a frozen four-class posterior over neutral, happy, angry and sad.

• The definitive evaluation uses identical stratified five-fold partitions repeated across three seeds for every model, producing fifteen paired observations per configuration. The leave-one-show-out protocol is a secondary domain-sensitivity stress test.

• The retained aggregate results support all numerical tables in the main manuscript and this supplement. The exact GPU identifier, fully frozen dependency lockfile, per-instance predictions, per-instance gate values and final-model error logs were not retained by the completed Google Colab environment and are therefore not presented as available artefacts.

• The experimental source code is documented in the doctoral thesis appendix and may be obtained from the corresponding author on reasonable request, subject to the original dataset terms. A future public release should include a version-pinned environment, fold indices, feature checksums, training logs, pooled out-of-fold predictions, per-instance gates, confusion matrices and a script reproducing the TOST sensitivity analysis.

S9. Supplementary artefact limitations and recommended future outputs

- Per-show feature diagnostics: future releases should compute audio duration, token length, lexical diversity, pitch range and projected-embedding variance by programme using training-independent procedures.
- Emotion subgroup analysis: valid analysis requires pooled out-of-fold predictions, fold-specific emotion outputs, fold-specific gate values, subgroup sizes and multiplicity control. These artefacts were not retained.
- Error analysis: future experiments should preserve 20–30 representative final-model errors, contexts, predicted probabilities, speaker/show identifiers and per-show confusion matrices.
- Hyperparameter selection: future work should use nested cross-validation or a prespecified validation set, with declared search spaces for learning rate, dropout, early-stopping patience, pooling method and classifier width.
- External validity: leave-one-speaker-out, balanced cross-show tests and controlled domain-adaptation experiments are required before speaker- or programme-general claims.

S10. Supplementary package index

Table S11. Files supplied with the article

File	Purpose	Recommended portal designation
PEFM_Supplementary_Material.docx		Supplementary Information
PEFM_Supplementary_Data.xlsx	Machine-readable corpus, settings, results, statistical tests, gates, LOSO and error-summary tables	Supplementary Data
PEFM_All_Figures_High_Resolution.docx		
PEFM_All_Tables_Editable.docx		
Figures/Fig1–Fig4.png and .eps		Figure files
README_SupplementARY_PACKAGE.txt		