Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Iyanuoluwa Fatoki
Lead City University
Victoria Bolanle Oyekunle
Lead City University
Emmanuel Adebowale Adediran
adediran.emmanuel@lcu.edu.ng

Lead City University

Research Article

Keywords: multimodal sarcasm detection, acoustic prosody, cross-modal incongruity, HuBERT, controlled comparison, cross-show generalisation

Posted Date: August 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10514016/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

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

Iyanuoluwa Fatoki[1] · Victoria Bolanle Oyekunle[1] · Emmanuel Adebowale Adediran[2]*

[1] Department of Computer Science, Lead City University, Ibadan, Nigeria
[2] Centre for Technology, Innovation and Advanced Research Development, Lead City University, Ibadan, Nigeria
*Corresponding author: Emmanuel Adebowale Adediran; adediran.emmanuel@lcu.edu.ng
Author e-mails: Fatoki.iyanuoluwa@lcu.edu.ng; bola.oyekunle@lcu.edu.ng; adediran.emmanuel@lcu.edu.ng

Abstract

Spoken sarcasm is often disambiguated by vocal delivery rather than by words alone. This study therefore examines whether acoustic-prosodic information, speech-derived emotion and explicit text-audio incongruity provide measurable value beyond strong text-only and text-plus-audio baselines. Six configurations were evaluated on the complete 690-utterance audio-bearing MUStARD corpus using stratified five-fold cross-validation repeated across three seeds. PEFM-CMAE combines contextual RoBERTa text features, mean-pooled HuBERT acoustic features and a speech-emotion posterior vector over neutral, happy, angry and sad. Its incongruity vector is learned from the absolute difference and element-wise product of projected text and audio representations, then supplied to a softmax modality gate and classifier. PEFM-CMAE produced the highest mean macro-F1 (0.7504), significantly exceeding text-only RoBERTa (0.6909; $p = 0.000002$) and the earlier PEFM-AE2 model (0.7325; $p = 0.001062$). Direct text-plus-audio fusion reached 0.7479; the 0.0025 difference was not statistically significant and was not demonstrably equivalent (95% CI -0.0106 to 0.0156; TOST $p = 0.120$ using a post-hoc ±0.010 bound). The emotion gate received a mean weight of 0.0331, indicating a supplementary rather than dominant role. Leave-one-show-out performance fell materially. The evidence therefore supports within-corpus acoustic complementarity but not robust generalisation to unseen shows.

Keywords: multimodal sarcasm detection; acoustic prosody; cross-modal incongruity; HuBERT; controlled comparison; cross-show generalisation

1 Introduction

Sarcasm is a pragmatic phenomenon in which a speaker's intended meaning diverges from the literal content of an utterance. In speech, the relevant evidence may be distributed across lexical choice, conversational context, timing, pitch, intensity and affective delivery. The established importance of prosody means that this paper does not treat the use of acoustic information as a new discovery. Its question is more specific: under a controlled protocol, how much do mean-pooled HuBERT acoustic features add to a contextual RoBERTa text model, and do speech-derived emotion and explicit text-audio incongruity add value beyond direct bimodal fusion [1-6]?

Phonetic and psycholinguistic work has documented changes in pitch range, duration, speaking rate and voice quality during ironic speech [5, 6]. Cognitive accounts further show that irony is interpreted within interpersonal and conversational context rather than from a single fixed marker [24]. These cues vary across speakers: some speakers exaggerate prosody, whereas others use flat or deadpan delivery. The computational task is therefore relational, involving the interaction between what is said, how it is said and the context in which it occurs. That relational premise motivates the explicit incongruity descriptor used here: the absolute difference and element-wise product of projected text and audio vectors encode disagreement and interaction rather than merely appending one modality to another.

Pretrained encoders provide strong representations for this comparison. RoBERTa captures contextual lexical information [7], while HuBERT learns phonetic and paralinguistic structure from raw waveforms [8]. Recent prosodic probing work confirms that self-supervised speech models retain structural prosodic information, although temporal aggregation choices affect what remains accessible [23]. The unresolved issue is not whether these encoders are useful in isolation, but whether a more elaborate fusion design produces reliable improvement over direct bimodal fusion.

Emotion is relevant for the same reason, but it may be redundant when inferred from text. The final model therefore uses a speech-derived posterior as a separate conditioning stream. The controlled comparison tests whether that coarse affective representation improves aggregate performance and how much weight the learned gate assigns to it; it does not assume in advance that emotion must be a dominant predictor.

The article reports the complete 690-utterance evaluation of PEFM-CMAE. The model combines contextual RoBERTa features, HuBERT acoustic-prosodic features, a four-class speech-emotion posterior, an explicit text-audio incongruity vector and a softmax modality gate. The analysis is framed primarily as a rigorous comparative and

interpretability study, with the architecture treated as an incremental, testable refinement rather than a claim of wholly new component technologies.

The contribution is fourfold. First, the study quantifies the acoustic contribution under identical partitions, showing that text-plus-audio fusion significantly improves upon text only. Second, it tests whether acoustic emotion, adaptive gating and explicit incongruity improve upon that strong simple baseline; the answer is deliberately reported as mixed rather than uniformly positive. Third, it makes the evaluated incongruity computation and modality allocation explicit, permitting inspection of what the final refinement adds. Fourth, it reports uncertainty, an equivalence sensitivity analysis and cross-show stress testing, thereby separating within-corpus performance from evidence of generalisation.

2 Related work and theoretical rationale

2.1 Evolution from text-only to spoken sarcasm recognition

Early computational sarcasm detection concentrated on punctuation, polarity reversal, interjections, lexical contrast and author context. Transformer models strengthened textual modelling, but transcripts cannot preserve vocal delivery. MUStARD paired sitcom utterances with context and audiovisual clips, enabling controlled multimodal investigation [1]. Subsequent work has explored multitask sentiment and emotion supervision [12], speaker or author representations [25], speech-emotion transfer [3, 4], hierarchical congruity [20], contrastive debiasing [21] and increasingly explicit cross-modal interaction.

Gao and colleagues established that speech alone contains sarcasm-relevant information and later combined speech, text, emotion and sentiment [3, 4]. Those studies leave a narrower comparative gap: when RoBERTa and HuBERT are held constant and evaluated on identical folds, is the gain mainly attributable to adding audio, or do an emotion posterior, adaptive gating and an incongruity block provide further reliable benefit? The present study addresses that question through six internally controlled configurations rather than treating the presence of audio itself as the novelty.

2.2 Prosodic, affective and cross-modal incongruity

Acoustic studies characterise irony through probabilistic changes in pitch height and range, rate, duration, loudness and voice quality [5, 6]. Cognitive-pragmatic work similarly treats irony as an interpretation constructed from context and speaker intention [24]. The same words can therefore be sincere in one setting and sarcastic in another, while the same sarcastic intention may be conveyed through exaggerated or attenuated prosody.

This relational view motivates explicit incongruity features. Prior multimodal work has modelled congruity through cross-attention, graph structures and contrastive objectives, mainly for image-text sarcasm [20, 21]. Recent published extensions include graph-attention modelling of emotional incongruity on MUStARD++ [26], LLM-assisted annotation of a larger speech-sarcasm corpus [27], and cue-learning over pretrained multimodal representations [28]. For spoken sarcasm, the corresponding theoretical prediction is bounded: when projected lexical and acoustic representations disagree, features derived from their absolute difference and element-wise interaction may assist classification. The present experiment tests whether that representation improves average classification; it does not directly label or isolate high-incongruity utterances.

2.3 Positioning and scope of the contribution

The components used here are established: RoBERTa, HuBERT, speech-emotion recognition, multilayer perceptrons and softmax gating. The methodological contribution is their controlled evaluation around a specific incongruity representation, coupled with paired partitions, modality-gate analysis and cross-show stress testing. This positioning avoids claiming that acoustic-prosodic sarcasm detection itself is new.

The paper also distinguishes three levels of evidence. A significant text-plus-audio gain supports acoustic complementarity. A further significant gain by PEFM-CMAE over PEFM-AE2 supports improvement over the earlier gated formulation. A non-significant 0.0025 difference from direct text-plus-audio fusion does not establish superiority, and equivalence requires a separate test. These distinctions govern the claims in the abstract, discussion and conclusion.

2.4 Research questions

The completed experiment addresses the following questions:

1. RQ1. Does the addition of HuBERT-derived acoustic information improve binary sarcasm classification over text-only RoBERTa?
2. RQ2. Does a speech-derived acoustic-emotion channel improve classification beyond text-plus-audio fusion?

3. RQ3. Does the PEFM-CMAE combination of cross-modal incongruity modelling and gated fusion improve upon shallower multimodal baselines, and what do the learned gate weights indicate about modality use?

3 Materials and methods

3.1 Corpus preparation

The final evaluation used the complete audio-bearing MUStARD corpus. The official annotation inventory was matched to the raw audiovisual clips; each audio track was extracted, converted to monophonic form and resampled to 16 kHz. All 690 annotated instances were retained as paired transcript-waveform observations. The executable preparation notebook records the source identifier for each clip so that the inventory can be audited [1].

The corpus is balanced globally, with 345 sarcastic and 345 non-sarcastic utterances. The source-show distribution is not balanced. Friends contributes 356 instances and The Big Bang Theory 280, whereas The Golden Girls contributes 40 and Sarcasmaholics Anonymous 14. The latter contains no non-sarcastic examples, and The Golden Girls contains only one. This structure is acceptable for stratified cross-validation but creates severe limitations for leave-one-show-out interpretation. The original distribution was retained rather than artificially resampled at show level.

A further cross-show diagnostic would ideally report, for each programme, audio-duration distribution, tokenised utterance length and dispersion of the cached RoBERTa and HuBERT embeddings. These per-show summaries were not preserved in the retained complete-corpus artefacts and cannot be reconstructed from the aggregate result files without rerunning feature extraction. They are therefore not estimated post hoc. Table 1 reports the verified sample and class-prior audit, while prospective cross-show experiments should store the mean, standard deviation and range of those feature statistics for every programme.

Table 1 Composition of the complete audio-bearing MUStARD corpus

Source show	Non-sarcastic	Sarcastic	Total
The Big Bang Theory	140	140	280
Friends	204	152	356
The Golden Girls	1	39	40
Sarcasmaholics Anonymous	0	14	14
Total	345	345	690

3.2 Input preparation and encoders

The textual input comprised the target utterance and up to three preceding context turns, separated with the RoBERTa separator token. A maximum of 192 subword tokens was selected from the corpus audit to accommodate this fixed context window while bounding memory use; longer sequences were right-truncated. RoBERTa-base supplied the final [CLS] representation. The model parameters of the encoder were frozen and its 768-dimensional outputs were cached before fold-level training.

Audio was resampled to 16 kHz and capped at 8 s (128,000 waveform samples). Clips exceeding the cap were right-truncated after resampling; the threshold retained complete audio for more than 95% of the corpus, while shorter clips were zero-padded at batch construction. HuBERT-base supplied frame-level 768-dimensional representations, which were aggregated by arithmetic mean pooling (mean pooling). Mean pooling introduced no trainable parameters and was chosen as a low-variance fixed-size summary for a small cached-feature experiment. A minimal single-score attention pool over 768-dimensional frames would add 769 parameters (a 768-dimensional scoring vector and bias) and could preserve temporal salience, but that alternative was not evaluated here.

A pretrained HuBERT-based speech-emotion recogniser generated a posterior vector over neutral, happy, angry and sad. The four probabilities formed h_e. RoBERTa, HuBERT and the emotion recogniser were frozen feature extractors; only the projection layers, incongruity network, modality gate and classifier were optimised. The encoded inputs are defined in Eqs. (1)-(3).

$$h_t = f_t(x_t)[\mathrm{CLS}] \in \mathbb{R}^{768} \quad (1)$$

$$h_a = T^{-1} \sum_{t=1}^{T} f_a(x_a)_t \in \mathbb{R}^{768} \quad (2)$$

$$h_e = P(e \mid x_a),\ e \in \{\text{neutral, happy, angry, sad}\} \quad (3)$$

3.3 PEFM-AE2 and PEFM-CMAE

PEFM-AE2 projects the text, acoustic and acoustic-emotion vectors into a shared 256-dimensional space: t = phi_t(h_t), a = phi_a(h_a) and e = phi_e(h_e), where each phi comprises a linear layer, layer normalisation and ReLU activation. PEFM-CMAE adds the explicit text-audio interaction shown in Fig. 1. The 512-dimensional descriptor [|t-a|; t⊙a] is mapped by an incongruity multilayer perceptron to a 256-dimensional vector i. The gate receives the 1,024-dimensional concatenation [t;a;e;i] and produces three modality weights. The 256-dimensional weighted fusion f is concatenated with i, producing a 512-dimensional classifier input. The final classifier has a 512-unit hidden layer and two output logits.

The evaluated computation is given in Eqs. (4)-(7). The gate notation is used consistently throughout: the incongruity vector i is part of the gate input, but the softmax produces only three weights for text, audio and emotion. The incongruity vector also reaches the classifier directly through z, so acoustic-text interaction can influence a prediction even when the scalar audio or emotion gate is small.

For PEFM-CMAE, the implemented computation is:

$$i = \varphi_i([|t - a| ; t \odot a]) \quad (4)$$

$$\alpha = \mathrm{softmax}(g([t ; a ; e ; i])) \quad (5)$$

$$f = \alpha_t t + \alpha_a a + \alpha_e e \quad (6)$$

$$z = [f ; i] \quad (7)$$

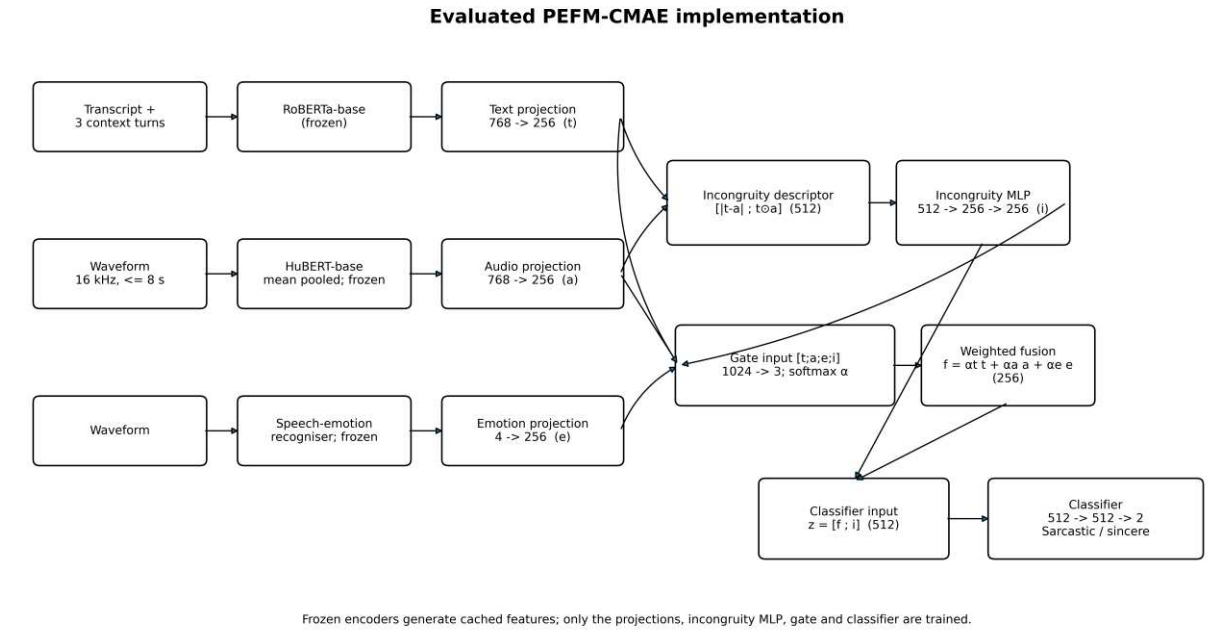


Fig. 1 Evaluated PEFM-CMAE architecture and tensor dimensions. Frozen encoders generate cached features; the incongruity vector is supplied to both the modality gate and classifier.

3.4 Comparative configurations

Six configurations were trained on identical partitions. Their sequence separates the value of each information source from the value of the incremental fusion mechanisms.

Table 2 Controlled model configurations

Configuration	Active information	Purpose
Text-only RoBERTa	Text + context	Lexical baseline
Audio-only HuBERT	Waveform	Acoustic-prosodic baseline
Text + audio	Text and waveform	Strong 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

3.5 Training and evaluation

The final evaluation used stratified five-fold cross-validation repeated across three random seeds, producing fifteen paired seed-fold partitions for every configuration. Macro-F1 was primary; accuracy, macro-precision and macro-recall were complementary. Paired t-tests compared fold-level macro-F1 on identical partitions. Ninety-five per cent confidence intervals were reconstructed from the paired mean differences and t statistics. For the marginal difference between PEFM-CMAE and text-plus-audio fusion, a post-hoc equivalence sensitivity analysis used two one-sided tests with a practical bound of ±0.010 macro-F1. This bound was not preregistered and is therefore reported as sensitivity analysis rather than confirmatory evidence.

Leave-one-show-out evaluation held out each programme in turn. It simultaneously changes speaker population, writing style and production conditions because speakers are largely nested within programmes. The protocol can therefore reveal domain sensitivity but cannot identify which source of shift caused a performance change. Two held-out programmes are also severely label-skewed, so leave-one-show-out results are treated as stress tests rather than a primary ranking.

The final cached-feature experiment used Python, PyTorch, Hugging Face Transformers, scikit-learn, NumPy and SciPy in a Google Colab GPU environment. The exact GPU model and dependency versions were not preserved in the retained aggregate outputs, because the notebook installed unpinned current packages and only printed the detected GPU at runtime. Rather than reconstructing these details, the omission is reported explicitly. A submission repository should include the executable notebook, a frozen requirements file, fold indices, feature checksums, per-fold predictions and package/hardware logs.

Table 3 Core experimental settings

Setting	Value
Text encoder	RoBERTa-base, frozen cached 768-d features
Acoustic encoder	facebook/hubert-base-ls960, frozen; mean-pooled 768-d features
Emotion stream	Four-class HuBERT-based posterior; frozen
Projected hidden dimension	Text/audio: 768 → 256; emotion: 4 → 256
Incongruity input	$[\lvert t - a \rvert ; t \odot a]$ (512 d) → MLP 512 → 256 → 256
Classifier	$z = [f ; i]$ (512 d) → hidden layer 512 → 2 logits; dropout applied
Optimiser	AdamW; weight decay 0.01
Fusion/head learning rate	1×10^{-3} in the cached-feature evaluation
Batch size	4
Maximum epochs	30; early stopping on validation macro-F1; patience 8
Early-stopping rationale	Heuristic fixed choice, not the result of a hyperparameter search; selected to avoid premature stopping under small-fold metric noise
Loss	Cross-entropy. Stratification produced 276 sarcastic and 276 non-sarcastic training items per primary fold, so recomputed inverse-frequency weights were 1.0 and 1.0 and had no numerical effect
Text/audio limits	192 subword tokens; 8 s at 16 kHz; longer inputs right-truncated
Primary protocol	5 folds × 3 seeds = 15 paired partitions
Stress test	Leave one source show out
Modality gate	$[t ; a ; e ; i]$ (1024 d) → linear 1024 → 3 → softmax

4 Results

4.1 Staged architectural development

The base feasibility phase used the 200-utterance audio-bearing subset. Audio-only modelling obtained macro-F1 = 0.696, closely matching text-only modelling at 0.692, while text-plus-audio reached 0.722. The placeholder emotion channel did not improve the full model, which obtained 0.695. This phase established that the speech signal was independently predictive and that an uninformative emotion input could be suppressed by the gate.

The second phase replaced the placeholder with a speech-derived posterior and redesigned the fusion mechanism. Across twenty-five stratified folds, PEFM-AE2 achieved macro-F1 = 0.7544 ± 0.0678, compared with

0.7475 ± 0.0561 for text plus HuBERT audio, 0.7443 ± 0.0605 for three-stream concatenation and 0.6426 ± 0.0703 for text only. The improvement over text only was significant ($p = 1.19 \times 10^{-6}$); the smaller margins over the two multimodal baselines were not significant. This result motivated the final complete-corpus evaluation rather than supporting an unqualified claim that gating alone was superior.

The definitive phase reconstructed all 690 audio-bearing utterances and added the explicit absolute-difference/product incongruity block. All main claims below are based on this shared complete-corpus evaluation.

Table 4 Summary of the three empirical phases

Phase	Corpus	Principal result
Base PEFM	200 utterances	Text 0.692; audio 0.696; text + audio 0.722; full PEFM 0.695
PEFM-AE2	200 utterances; 25 seed-folds	PEFM-AE2 0.7544 ± 0.0678; strongest mean, but not significantly above strong multimodal baselines
PEFM-CMAE	690 utterances; 15 seed-folds	PEFM-CMAE 0.7504 ± 0.0334; significantly above text only and PEFM-AE2

4.2 Complete-corpus stratified results

PEFM-CMAE recorded the highest mean of the six configurations (macro-F1 0.7504 ± 0.0334), followed closely by text plus audio (0.7479 ± 0.0400) and three-stream concatenation (0.7436 ± 0.0329). Audio-only HuBERT obtained 0.7291 ± 0.0406, exceeding text-only RoBERTa at 0.6909 ± 0.0346. PEFM-AE2 obtained 0.7325 ± 0.0344. The ordering shows that acoustic information is valuable, whereas additional emotion and gating do not automatically improve upon direct bimodal fusion.

Table 5 Complete-corpus stratified performance (mean ± SD over 15 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

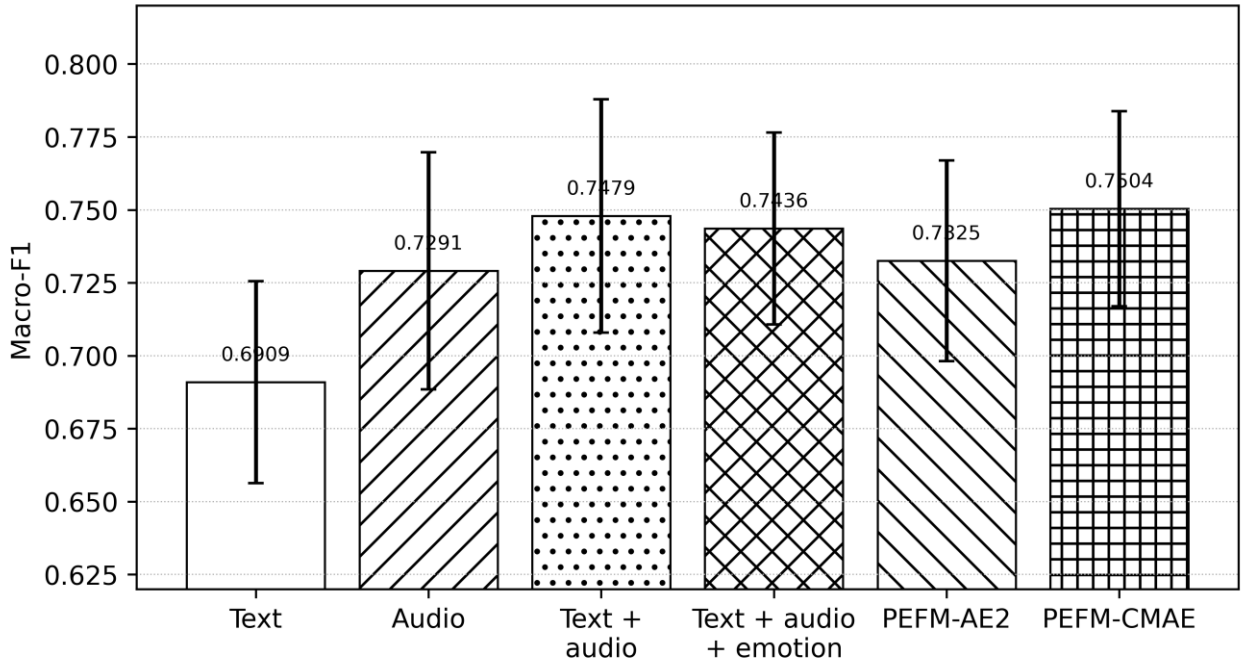


Fig. 2 Mean macro-F1 and standard deviation for the six complete-corpus configurations across fifteen paired seed-fold partitions

4.3 Paired statistical comparisons

PEFM-CMAE exceeded text-only RoBERTa by 0.0595 macro-F1 (t = 7.8531, p = 0.000002; 95% CI 0.0432 to 0.0758; Cohen's d = 2.03). Text plus audio also exceeded text only by 0.0570 (p = 0.000021; 95% CI 0.0374 to 0.0766; d = 1.61). These large effects support the value of acoustic-prosodic information under the common protocol.

PEFM-CMAE improved upon PEFM-AE2 by 0.0179 (p = 0.001062; 95% CI 0.0086 to 0.0272; d = 1.06). By contrast, its 0.0025 margin over text plus audio was not significant (p = 0.688882; 95% CI -0.0106 to 0.0156; d = 0.11). The post-hoc TOST sensitivity analysis used ±0.010 macro-F1 as the smallest effect of practical interest: a difference below one absolute macro-F1 point was judged too small to justify the added fusion complexity and was also smaller than the observed between-partition variability. TOST p = 0.120 did not establish equivalence. This bound was not preregistered and is therefore interpreted as sensitivity analysis rather than confirmatory evidence.

Table 6 Paired macro-F1 comparisons on identical seed-fold 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

4.4 Learned modality gates

Both gated models retained a dominant text pathway. PEFM-AE2 assigned mean weights of 0.6024 to text, 0.3546 to audio and 0.0430 to acoustic emotion; PEFM-CMAE assigned 0.7312, 0.2357 and 0.0331. The final three-stream concatenation (0.7436) and PEFM-AE2 (0.7325) did not exceed direct text-plus-audio fusion (0.7479). The complete-corpus results therefore provide no aggregate evidence that the four-class emotion posterior adds a reliable performance gain.

The 0.0331 mean emotion weight is best interpreted as underutilisation or a supplementary conditioning role, not as evidence of strong independent contribution. Per-instance gate values were not retained in the supplied aggregate artefacts, so the study cannot identify utterances with unusually high emotion reliance or test deadpan and neutral-valence subsets retrospectively. Future runs should save instance-level gates and compare discrete posteriors with continuous valence-arousal features, larger speech-emotion embeddings, and utterance-only versus context-conditioned emotion estimation. Gate weights remain descriptive and are not causal explanations [14].

Table 7 Learned modality gate weights on the complete corpus

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

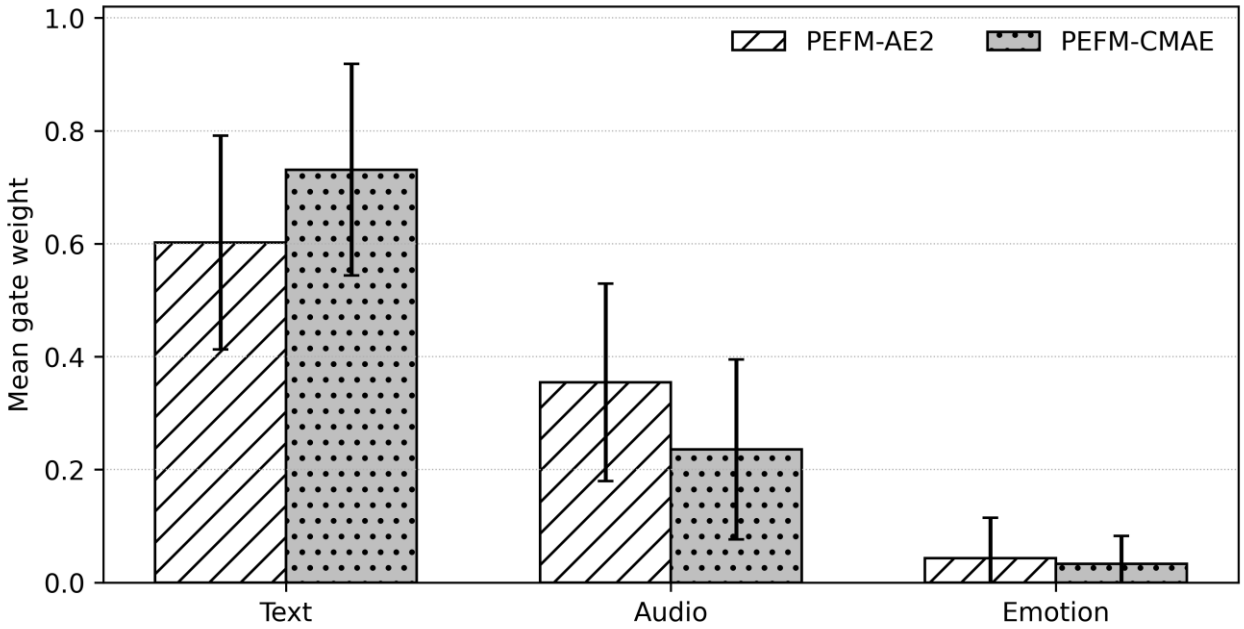


Fig. 3 Mean learned modality gate weights for PEFM-AE2 and PEFM-CMAE across fifteen seed-fold partitions. Error bars denote standard deviation

4.5 Leave-one-show-out stress test

Cross-show performance was materially below the stratified mean. PEFM-CMAE fell from 0.7504 to 0.6713 on Friends and 0.6426 on The Big Bang Theory; PEFM-AE2 was marginally stronger on Friends, while PEFM-CMAE was strongest on The Big Bang Theory. The label priors also differ sharply by programme: sarcasm accounts for 42.7% of Friends, 50.0% of The Big Bang Theory, 97.5% of The Golden Girls and 100% of Sarcasmaholics Anonymous. The two small one-class or nearly one-class folds therefore cannot support stable model ranking.

The Golden Girls has 40 observations but only one non-sarcastic item, while Sarcasmaholics Anonymous contains 14 sarcastic items and no non-sarcastic item. In addition, named characters are largely programme-specific. Leave-one-show-out evaluation consequently removes the principal speaker pool together with the script style, conversational relationships and recording domain. It is a joint speaker-domain shift rather than a clean leave-one-speaker-out test, and the observed decline cannot be attributed to any single source without further experiments.

Table 8 Leave-one-show-out macro-F1 by held-out programme

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

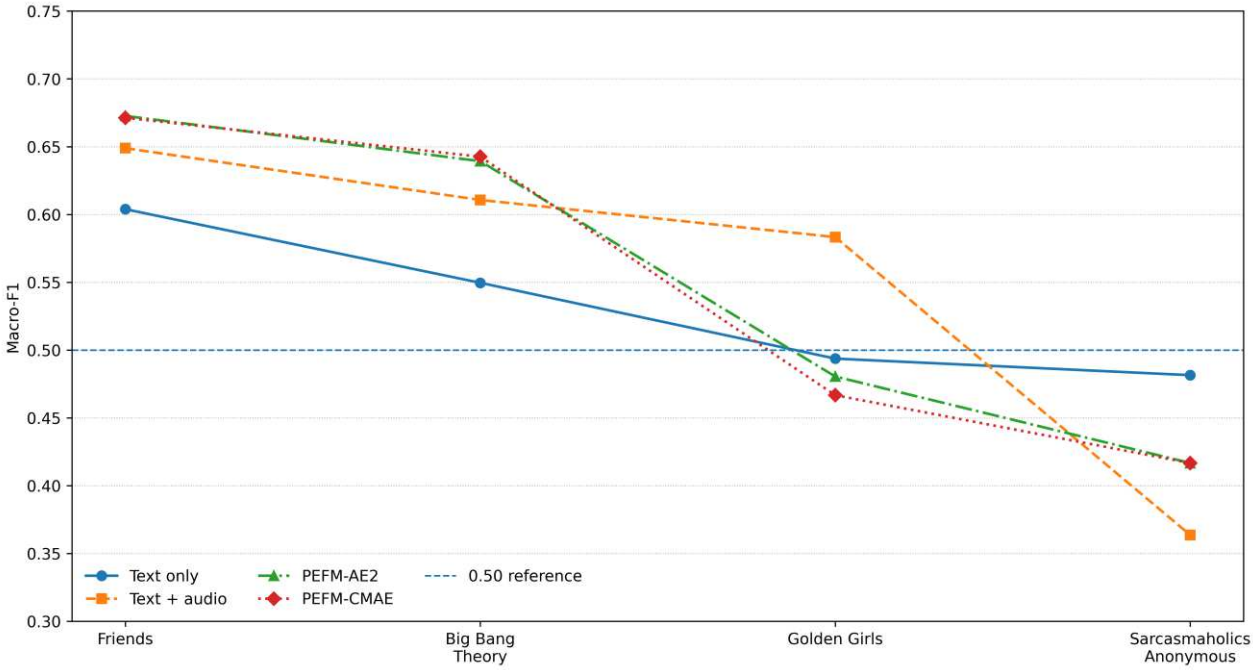


Fig. 4 Leave-one-show-out macro-F1 by held-out source show. The horizontal line marks 0.50; results for The Golden Girls and Sarcasmaholics Anonymous are unstable because of small and severely label-skewed folds.

4.6 Contextual comparison with reported MUStARD results

Table 9 places the internally controlled findings beside representative published MUStARD results. The values are contextual rather than a leaderboard because speaker-dependent, speaker-independent and repeated-stratified protocols are different estimands, and prior papers do not always report macro-F1. Gao et al. [4] reported F1 = 0.810 using attention-based text-audio fusion with emotion and sentiment information under a different speaker-independent protocol; the higher value may reflect the split, metric variant, augmentation or modelling choices and is not directly comparable with the repeated macro-F1 reported here. Macro-F1 gives equal weight to both classes and may be lower than micro- or weighted F1 on imbalanced partitions. The older SVM results remain useful only as historical anchors for the original MUStARD benchmark.

Table 9 Indicative comparison with published MUStARD results; protocols and metrics are not directly comparable

Study/model	Modalities	Protocol	Reported metric	Value
Castro et al. [1], SVM	Text	Speaker-independent	F1	0.609
Castro et al. [1], SVM	Text + audio	Speaker-independent	F1	0.631
Castro et al. [1], SVM	Text + audio + video	Speaker-independent	F1	0.643
Chauhan et al. [12], multitask	Text + audio + video	Speaker-independent	Accuracy	0.7694
Gao et al. [4], 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. [26]	Text + audio + video	MUStARD++; study-specific protocol	Macro-F1	0.7496

Note. Published values are included for contextual orientation only. Speaker-independent, speaker-dependent and repeated stratified cross-validation protocols estimate different generalisation targets; reported F1 variants, class distributions, modalities and augmentation procedures also differ. The rows must not be interpreted as a direct model ranking.

5 Discussion

5.1 Value of the acoustic-prosodic channel

The most robust finding is the contribution of speech-derived information. Audio-only HuBERT exceeded text-only RoBERTa by 0.0382 macro-F1, and direct text-plus-audio fusion exceeded text only by 0.0570 with a significant paired difference. This supports acoustic complementarity, consistent with linguistic evidence that sarcastic intent is often carried by delivery [5, 6]. It does not establish that the specific PEFM refinements are necessary for obtaining the multimodal gain.

Mean pooling provides a useful low-variance summary for a small dataset and frozen-feature experiment, but it removes temporal localisation. A minimal scalar attention pool would add only 769 trainable parameters, so parameter count alone is not a decisive objection; the stronger concern is that any learned temporal scorer must be estimated from a small corpus and could overfit programme- or speaker-specific timing patterns. Attention pooling, recurrent aggregation and temporally resolved HuBERT layers should therefore be compared under the same folds rather than assumed to improve performance.

5.2 What the emotion channel contributes

The emotion findings are weaker than the acoustic findings. A speech-derived posterior made the affective pathway operational, but in the final common evaluation neither three-stream concatenation nor PEFM-AE2 surpassed direct text-plus-audio fusion. PEFM-CMAE assigned the emotion stream a mean gate weight of 0.0331. The aggregate evidence therefore supports only a supplementary role for the coarse neutral/happy/angry/sad posterior vector.

A coarse posterior may still help particular utterances, but the retained artefacts do not contain the per-instance emotion categories, gate values and pooled out-of-fold predictions required for valid subgroup or correlation analysis. The manuscript therefore does not claim selective benefit on angry, neutral, deadpan or high-conflict cases. A definitive follow-up should stratify by predicted emotion, lexical sentiment, audio duration, show identity and text-audio uncertainty, and compare the four-class posterior with continuous valence-arousal features and richer emotion embeddings. Such analyses should use pooled out-of-fold predictions and fold-specific gate values, generate emotion predictions without access to held-out labels, report subgroup sizes, and control the family-wise error or false-discovery rate when multiple subgroups are tested.

5.3 Contribution of cross-modal incongruity modelling

PEFM-CMAE significantly improved upon PEFM-AE2, showing that the implemented absolute-difference/product interaction corrected a weakness of the earlier gate. However, the margin over direct text-plus-audio fusion was small, its confidence interval included effects in both directions and equivalence was not established. The contribution is consequently an interpretable comparative result and an explicit incongruity implementation, not decisive superiority over a simple bimodal system.

The theoretical hypothesis is correspondingly specific: when lexical and acoustic representations conflict, $|t - a|$ should increase, while $t \odot a$ records aligned or opposing feature interactions after projection. The MLP may use these signals to alter modality weights or classification. The current study tests the aggregate consequence of this formulation but does not directly verify the predicted subgroup effect. A stronger theoretical test requires labelled or operationalised lexical-affective mismatch strata.

5.4 Generalisation and error sources

The cross-show decline is a critical external-validity limitation. Speakers are largely nested within shows, so speaker identity is confounded with script style, recurring relationships, label prior and production conditions. Available corpus composition already shows substantial domain shift: the held-out programmes differ in size and sarcasm prevalence, and the two smallest folds are nearly or completely single-class. The retained experiment did not preserve per-show audio-duration, token-length, lexical-diversity, pitch-range, embedding-variance or final-model confusion summaries. Consequently, the article does not invent numerical post-hoc feature distributions, and the absence of these diagnostics is treated as a limitation of the present cross-show analysis.

A definitive diagnosis requires leave-one-speaker-out testing, balanced cross-show evaluation, pooled out-of-fold predictions and controlled adaptation experiments. Appropriate remedies include show-balanced sampling or loss reweighting, domain-adversarial training with a show classifier and gradient reversal, CORAL or maximum-mean-discrepancy alignment of projected modality features, speaker-adversarial objectives, leave-domain-out episodic training and few-shot calibration on a small labelled sample from a new programme. These methods should be compared with a no-adaptation baseline so that any recovery is not attributed merely to extra target-domain labels.

5.5 Available exploratory error evidence

The retained thesis contains a manual review of 11 errors from one fold of the earlier 200-utterance phase, not from the final 690-utterance PEFM-CMAE evaluation. It is retained only to identify plausible failure modes: context dependency accounted for 4 errors (36.4%), absence of lexical cues for 3 (27.3%), prosodic ambiguity for 3 (27.3%) and speaker drift for 1 (9.1%). These percentages are not presented as the final model's error distribution.

Table 10 Exploratory error categories from one earlier-phase fold; not a final-model error distribution

Error category	Count	Share (%)	Interpretation
Context dependency	4	36.4	Evidence beyond the fixed context window

Error category	Count	Share (%)	Interpretation
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

The final-model artefacts available for this article do not include pooled instance-level predictions, contexts and gate values, so a 20-30 case final-model audit, per-show confusion matrices and representative prediction examples cannot be reconstructed reliably. Future evaluation must save these artefacts prospectively and report error distributions by show, speaker, context length, audio duration, lexical-affective mismatch and predicted emotion. This limitation is stated explicitly rather than substituting examples from a different experimental phase.

5.6 Future work

The next evaluation should prioritise three linked questions. First, external validity should be separated into speaker and programme effects through leave-one-speaker-out and balanced leave-one-show-out protocols, followed by the adaptation methods described above. Second, temporal and interaction mechanisms should be ablated directly: mean pooling versus scalar attention, recurrent aggregation and cross-attention; absolute-difference/product incongruity versus bilinear, tensor-fusion and cross-attention alternatives. Third, the affective pathway should be tested with continuous valence-arousal, richer embeddings and subset analyses on high text-audio disagreement. A proper hyperparameter study should use nested cross-validation or a prespecified validation partition for learning rate, dropout, early-stopping patience, pooling choice and classifier width, with the search space and selection criterion declared before final testing. All experiments should preserve pooled out-of-fold predictions, per-instance gates, confusion matrices and environment logs so that error, subgroup and equivalence analyses can be reproduced.

6 Limitations and reproducibility requirements

The study has six main limitations. MUStARD represents scripted English-language American sitcom dialogue rather than spontaneous, multilingual or code-switched sarcasm. The show distribution is uneven and speaker identity is confounded with source programme. The model excludes the visual modality. The four-class emotion posterior is coarse and receives little aggregate gate weight. Mean-pooled audio discards temporal prosodic structure. Finally, the retained final experiment lacks per-instance gates, pooled out-of-fold prediction files, a leave-one-speaker-out experiment and an exact GPU/dependency log.

Published MUStARD results use different speaker protocols, modalities and augmentation, so Table 9 is contextual rather than comparative evidence. The strongest claims rest on the six configurations evaluated on identical partitions. Additional literature on hierarchical congruity, cross-modal graphs, explanation and out-of-distribution debiasing [20–22] further underscores that robust generalisation requires more than a small within-corpus mean advantage.

The original MUStARD annotations and media access instructions are available from the dataset providers cited in [1]. Raw copyrighted clips are not redistributed. The authors' preparation logic, model specification, fold definitions, aggregate results and supporting code are available from the corresponding author on reasonable request, subject to the original dataset terms. A reproducibility release should include exact package versions, training logs, GPU metadata and a script reproducing the TOST analysis; those records were not all retained by the Google Colab environment used for the completed evaluation.

7 Conclusion

This controlled study shows that acoustic-prosodic information materially improves contextual text-based sarcasm detection on the complete MUStARD corpus. PEFM-CMAE attained the highest mean macro-F1 (0.7504), significantly exceeding text-only RoBERTa and PEFM-AE2. Its 0.0025 difference from direct text-plus-audio fusion was not significant and was not demonstrably equivalent, while the four-class emotion stream received little aggregate gate weight. The appropriate conclusion is therefore not that a complex architecture decisively surpasses direct fusion. Rather, speech provides reliable complementary evidence within MUStARD, the explicit incongruity design improves the earlier gated formulation, and robust transfer to unseen shows remains unresolved.

Funding

The authors received no specific funding for this work.

References

[1] Castro S, Hazarika D, Pérez-Rosas V, Zimmermann R, Mihalcea R, Poria S (2019) Towards multimodal sarcasm detection (an obviously perfect paper). In: Proceedings of ACL, pp 4619–4629. https://doi.org/10.18653/v1/P19-1455

[2] Potamias RA, Siolas G, Stafylopatis AG (2020) A transformer-based approach to irony and sarcasm detection. Neural Comput Appl 32:17309–17320. https://doi.org/10.1007/s00521-020-05102-3

[3] Gao X, Nayak S, Coler M (2022) Deep CNN-based inductive transfer learning for sarcasm detection in speech. In: Proceedings of Interspeech 2022, pp 2323–2327. https://doi.org/10.21437/Interspeech.2022-11323

[4] Gao X, Nayak S, Coler M (2024) Improving sarcasm detection from speech and text through attention-based fusion exploiting the interplay of emotions and sentiments. Proc Meet Acoust 54(1). https://doi.org/10.1121/2.0001918

[5] Anolli L, Ciceri R, Infantino MG (2000) Irony as a game of implicitness: acoustic profiles of ironic communication. J Psycholinguist Res 29(3):275–311. https://doi.org/10.1023/A:1005100221723

[6] Cheang HS, Pell MD (2008) The sound of sarcasm. Speech Commun 50(5):366–381. https://doi.org/10.1016/j.specom.2007.11.003

[7] Liu Y, Ott M, Goyal N et al (2019) RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692

[8] Hsu W-N, Bolte B, Tsai Y-HH, Lakhotia K, Salakhutdinov R, Mohamed A (2021) HuBERT: self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans Audio Speech Lang Process 29:3451–3460. https://doi.org/10.1109/TASLP.2021.3122291

[9] Arevalo J, Solorio T, Montes-y-Gómez M, González FA (2017) Gated multimodal units for information fusion. In: ICLR Workshop.

[10] Tsai Y-HH, Bai S, Liang P et al (2019) Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of ACL, pp 6558–6569. https://doi.org/10.18653/v1/P19-1656

[11] Hazarika D, Zimmermann R, Poria S (2020) MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In: Proceedings of ACM Multimedia, pp 1122–1131. https://doi.org/10.1145/3394171.3413678

[12] Chauhan DS, Dhanush SR, Ekbal A, Bhattacharyya P (2020) Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In: Proceedings of ACL, pp 4351–4360. https://doi.org/10.18653/v1/2020.acl-main.401

[13] Zadeh A, Chen M, Poria S, Cambria E, Morency L-P (2017) Tensor fusion network for multimodal sentiment analysis. In: Proceedings of EMNLP, pp 1103–1114. https://doi.org/10.18653/v1/D17-1115

[14] Jain S, Wallace BC (2019) Attention is not explanation. In: Proceedings of NAACL-HLT, pp 3543–3556. https://doi.org/10.18653/v1/N19-1357

[15] Poria S, Hazarika D, Majumder N et al (2019) MELD: a multimodal multi-party dataset for emotion recognition in conversations. In: Proceedings of ACL, pp 527–536. https://doi.org/10.18653/v1/P19-1050

[16] Busso C, Bulut M, Lee C-C et al (2008) IEMOCAP: interactive emotional dyadic motion capture database. Lang Resour Eval 42:335–359. https://doi.org/10.1007/s10579-008-9076-6

[17] Ray A, Mishra S, Nunna A, Bhattacharyya P (2022) A multimodal corpus for emotion recognition in sarcasm. In: Proceedings of LREC.

[18] Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. In: Proceedings of ICLR.

[19] Eyben F, Scherer KR, Schuller BW et al (2016) The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. IEEE Trans Affect Comput 7(2):190–202. https://doi.org/10.1109/TAFFC.2015.2457417

[20] Liu H, Wang W, Li H (2022) Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. In: Proceedings of EMNLP, pp 4995–5006. https://doi.org/10.18653/v1/2022.emnlp-main.333

[21] Jia M, Xie C, Jing L (2024) Debiasing multimodal sarcasm detection with contrastive learning. Proc AAAI Conf Artif Intell 38(16):18354–18362. https://doi.org/10.1609/aaai.v38i16.29795

[22] Kumar S, Mondal I, Akhtar MS, Chakraborty T (2023) Explaining (sarcastic) utterances to enhance affect understanding in multimodal dialogues. Proc AAAI Conf Artif Intell 37(11). https://doi.org/10.1609/aaai.v37i11.26526

[23] de Seyssel M, Lavechin M, Titeux H et al (2023) ProsAudit, a prosodic benchmark for self-supervised speech models. In: Proceedings of Interspeech, pp 2963–2967.

[24] Gibbs RW (2000) Irony in talk among friends. Metaphor Symb 15(1–2):5–27. https://doi.org/10.1080/10926488.2000.9678862

[25] Kolchinski YA, Potts C (2018) Representing social media users for sarcasm detection. In: Proceedings of EMNLP, pp 1115–1121. https://doi.org/10.18653/v1/D18-1140

[26] Raghuvanshi D, Gao X, Li Z, Bansal S, Coler M, Kumar N, Nayak S (2025) Intra-modal relation and emotional incongruity learning using graph attention networks for multimodal sarcasm detection. In: Proceedings of ICASSP 2025. IEEE. https://doi.org/10.1109/ICASSP49660.2025.10887864

[27] Li Z, Zhang Y, Gao X, Nayak S, Coler M (2025) Leveraging large language models for sarcastic speech annotation in sarcasm detection. In: Proceedings of Interspeech 2025, pp 3973-3977. https://doi.org/10.21437/Interspeech.2025-2074

[28] Lu M, Dong Z, Guo Z, Zhang X, Lu X, Wang T, Zhang L (2025) A multi-modal sarcasm detection model based on cue learning. Sci Rep 15:10261. https://doi.org/10.1038/s41598-025-94266-w

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- PEFMSupplementaryData.xlsx
- PEFMSupplementaryMaterial.pdf