Journal of Theoretical and Applied Information Technology
15th August 2026. Vol.104. No.15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CROSS-DOMAIN EMOTION DETECTION USING LIGHTWEIGHT LEXICON-GUIDED FEATURE FUSION: A THREE-DATASET STUDY WITH EXPLAINABILITY AND CONTRASTIVE ALIGNMENT ANALYSIS

K. HEMAKIRTHIGA[1],J. ARUNADEVI[2]✉

[1]Ph.D. Research Scholar (Part-Time)
PG & Research Department of Computer Science,
Raja Doraisingam Govt. Arts College, Sivaganga, Tamilnadu, India
Affiliated to Alagappa University.

[2] ✉ Assistant Professor
PG & Research Department of Computer Science,
Raja Doraisingam Govt. Arts College, Sivaganga, Tamilnadu, India
Affiliated to Alagappa University.
✉ Corresponding author: J.Arunadevi, E-mail: arunardm2015@gmail.com

ABSTRACT

Emotion detection models trained on one social media platform generalize poorly to another, yet this cross-domain generalization problem receives comparatively little attention relative to within-dataset benchmarking. We study it directly: we propose a lightweight, fully CPU-trainable architecture that fuses TF-IDF/SVD text embeddings with NRC Emotion Lexicon association features through a learned gate, and use it as the common architecture for a systematic three-dataset cross-domain generalization study spanning tweets (SemEval-2018 E-c), Reddit comments (GoEmotions), and personal narrative essays (ISEAR). Our central, statistically robust finding is that simple joint multi-domain training closes most of the cross-domain generalization gap relative to zero-shot single-domain transfer (bootstrap 95% CI for the Macro-F1 gain on GoEmotions: [0.095, 0.129]). As one specific, rigorously tested question within this study, we ask whether supervised contrastive alignment can improve on this strong, simple joint-training baseline; despite a dedicated hyperparameter search, a memory-bank negative pool, a stricter Jaccard-based positive-pair criterion, higher representation capacity, and evaluation across all three test domains, we find no configuration that statistically significantly outperforms naive pooling a considered negative result for this specific mechanism, not a limitation of the cross-domain generalization finding itself, which stands independently. We additionally report within-domain results against five baselines (our fusion model is competitive with, but does not exceed, a strong TF-IDF+SVM baseline we position the architecture explicitly as a vehicle for generalization and interpretability rather than raw within-domain accuracy), a real measured computational-efficiency comparison, an ablation testing whether the learned gate's near-constant empirical behavior (std = 0.012) means its adaptivity is dispensable (it is not: a fixed-scalar-gate ablation performs measurably worse, p = 0.0625), a multi-head gating variant, model-intrinsic explainability (gate reliance, SHAP over the interpretable lexicon channel), and a systematic label-confusion error analysis.

Keywords: *Cross-Domain Generalization; Emotion Detection; Social Media NLP; Lexicon Feature Fusion; Supervised Contrastive Learning; Multi-Label Classification; Explainable AI; Error Analysis*

1. INTRODUCTION

Automatic emotion detection in short, informal social media text underpins applications ranging from mental health monitoring to crisis response and brand analytics. Unlike coarse sentiment polarity, fine-grained emotion detection must handle multi-label outputs a single utterance can simultaneously express joy and surprise severe class imbalance, and, the central concern of this paper, a marked lack of robustness when a model trained on one platform's data (e.g., Twitter) is deployed on another (e.g., Reddit), where topical

content, register, and even the emotion taxonomy itself differ.

Two benchmark resources anchor most recent work. SemEval-2018 Task 1 (E-c) provides a multi-label emotion classification task over English tweets annotated for eleven emotions [1]. GoEmotions is a much larger corpus of Reddit comments annotated for twenty-seven fine-grained emotions plus neutral [2]. Because these corpora differ in domain, taxonomy granularity, and annotation protocol, models optimized on one transfer poorly to the other a problem that receives comparatively little attention relative to within-dataset benchmarking.

Problem Selection and Literature Screening

We chose the cross-domain generalization task because real-world deployment of emotion detection systems does not typically remain confined to a single platform, while most recent studies still report only within-dataset metrics. To narrow the scope of the literature, the query combinations "cross-domain / cross-corpus emotion detection", "lexicon feature fusion emotion", and "supervised contrastive learning multi-label / domain generalization" were used, restricted to peer-reviewed venues from 2018–2025 (ACL, EMNLP, COLING, SemEval, and major journals). Systems that either report out-of-domain results or explicitly combine lexicon and learned representations were given priority. The most directly comparable systems are Kim et al. [3], who combined lexicon features with an attention-based convolutional network on the SemEval task, and Zanwar et al. [4], who tested transformer representations combined with psycholinguistic features across several corpora. The closest methodological precedent for memory-based supervised contrastive domain generalization is provided by Tan et al. [5], who studied sentiment and rumour detection rather than multi-label emotion classification.

Differentiation from Prior Work

Kim et al. [3] focus on a single dataset and do not investigate cross-domain transfer. Zanwar et al. [4] do evaluate cross-corpus transfer, but they rely on a fine-tuned transformer backbone and report mostly positive transfer results; they neither isolate the effect of simple pooled training nor test a fully CPU-trainable, non-pretrained architecture. Tan et al. [5] show that memory-based contrastive alignment can benefit domain generalization when pretrained language-model encoders are used and when the tasks are different classification problems. The practical constraint addressed in the present study is different: no pretrained transformer downloads are required, and the entire pipeline remains trainable on a free-tier CPU. The scientific question is also different: once a strong, simple joint-training baseline has been established, does supervised contrastive alignment still provide benefit under a lightweight feature regime? We examine this question from multiple angles that have not previously been investigated for this class of architecture.

Research Objectives and Scope

This study pursues three specific objectives:

(O1) Quantify the cross-domain generalization gap for multi-label emotion detection under a shared-emotion subset protocol across three structurally distinct English corpora (SemEval-2018 E-c tweets, GoEmotionsReddit comments, and ISEAR narrative essays), and measure how much of that gap is closed by simple joint multi-domain training relative to zero-shot single-domain transfer.

(O2) Rigorously test whether supervised contrastive alignment (in-batch and memory-bank variants, hyperparameter search, capacity scaling, and a stricter Jaccard positive-pair criterion) can improve upon the joint-training baseline obtained in O1.

(O3) Characterise a lightweight, fully CPU-trainable gated fusion of TF-IDF/SVD text features and NRC Emotion Lexicon features with respect to within-domain competitiveness, computational cost, gate adaptivity, model-intrinsic explainability, and systematic error patterns.

What This Study Does Not Include

We deliberately exclude (a) fine-tuned pretrained transformer baselines, because the paper's contribution targets the opposite end of the resource spectrum; (b) non-English languages; (c) exhaustive neural architecture search or novel building blocks beyond the gated fusion mechanism; (d) training regimes substantially longer than eight epochs; and (e) alternative contrastive formulations beyond the supervised contrastive loss examined here. These exclusions

are stated so that the scope of the claims remains transparent.

This paper makes three contributions, centered on the cross-domain generalization question. First, we conduct a systematic three-dataset cross-domain generalization study spanning tweets, Reddit comments, and personal narrative essays (SemEval-2018 E-c, GoEmotions, and ISEAR) and show that simple joint multi-domain training closes most of the generalization gap relative to zero-shot single-domain transfer, a large and statistically robust effect. Second, as one specific, carefully investigated question within this study, we rigorously test whether supervised contrastive alignment can improve on this strong joint-training baseline not with a single implementation, but across a hyperparameter search, a stricter positive-pair definition, and a memory-bank negative-sampling design and report a considered negative result for this particular mechanism, which we present as an investigated component of the study rather than as the paper's central claim. Third, we support both findings with a lightweight fusion architecture (TF-IDF/SVD text representation plus NRC Emotion Lexicon features via a learned gate) that requires no pretrained transformer downloads, a real measured computational-efficiency comparison, a targeted ablation testing whether the gate's near-constant empirical behavior means its adaptivity is dispensable, model-intrinsic explainability, and a systematic error analysis.

We report our findings candidly, including the contrastive-alignment negative result reached only after genuine, multi-angle effort to overturn it in-batch, memory-bank, higher-capacity, hyperparameter-searched, and with a stricter positive-pair criterion, across three held-out test domains. This negative result does not diminish the paper's central finding: naive pooled training across domains, using nothing more than the lightweight fusion architecture itself, already delivers substantial, statistically robust cross-domain generalization gains, and is a strong, cheap-to-implement default for anyone deploying emotion detection across more than one social media platform.

2. RELATED WORK

2.1 Emotion Detection Benchmarks and Methods

SemEval-2018 Task 1 introduced a suite of tasks for affect in tweets, including the E-c multi-label emotion classification subtask used here, establishing a widely used English, Arabic, and Spanish benchmark across eleven emotion categories [1]. GoEmotions substantially extended taxonomy granularity, releasing 58k human-annotated Reddit comments across twenty-seven categories plus neutral, together with a fine-tuned BERT baseline [2]. Surveys of transformer-based emotion detection document the general shift from lexicon- and feature-based classifiers toward fine-tuned pretrained language models, while noting that most reported gains are measured within a single benchmark rather than across benchmarks [6].

2.2 Emotion Lexicons and Feature-Based Approaches

The NRC Emotion Lexicon associates roughly fourteen thousand English words and phrases with eight basic emotions and two polarities via large-scale crowdsourced annotation, and remains a widely used affective resource for feature engineering and interpretability [7]. Lexicon features have previously been combined with neural architectures for the SemEval-2018 E-c task, for example via attention-based convolutional networks that incorporate NRC associations alongside learned representations [3]. Our gating mechanism is conceptually related but differs by learning an example-dependent mixing weight between the lexicon and text streams, which we exploit directly for explainability.

2.3 Cross-Domain Generalization and Contrastive Alignment

Cross-domain transfer for emotion recognition has been studied extensively in speech and vision, including joint deep transfer learning across emotion corpora [8], but comparatively less for short-text social media emotion classification specifically across differing taxonomies. Most directly relevant to our study, Zanwar et al. [4] address cross-corpus generalizability for text-based emotion detection by combining transformer representations with psycholinguistic features, evaluating within-domain on GoEmotions and ISEAR the same two of our three datasets before testing out-of-domain transfer on six further

corpora from the Unified Emotion Dataset [9], and report that hybridlexical/psycholinguistic-plus-neural representations generalize better across domains than a standard transformer-based approach alone. In the broader text classification literature, domain generalization has also been approached via memory-based supervised contrastive learning that pulls same-class examples from different source domains together in embedding space while pushing different classes apart, evaluated on sentiment and rumour-detection datasets [5]. Our alignment term follows the same family of supervised contrastive objectives [10], adapted to the multi-label emotion setting by defining positive pairs via label-set overlap on a shared Ekman subset, and applied to a from-scratch, non-transformer encoder rather than for fine-tuning a pretrained language model as in prior NLP applications of supervised contrastive objectives [11]. Our memory-bank negative-sampling variant (Section 6.3) is directly inspired by momentum contrast, which demonstrated that a large, consistent dictionary of negatives maintained as a queue decoupled from the current mini-batch substantially improves contrastive representation learning in vision [12]; to our knowledge, applying this specific mechanism to cross-domain multi-label emotion alignment has not been previously reported, and we use it here as a rigorous test of whether the negative-pool-size explanation accounts for weak contrastive gains, rather than as a claimed positive result.

2.4 Explainability in Text Classification

Post-hoc feature attribution methods such as SHAP provide additive, game-theoretically motivated explanations of individual predictions and are widely used for auditing text classifiers [13]. We apply SHAP specifically to the compact, human-interpretable ten-dimensional NRC lexicon feature space rather than to opaque latent text embeddings, yielding explanations expressed directly in terms of named emotion and polarity categories.

2.5 Statistical Practice in NLP Evaluation

Reporting a single train/test split result is known to overstate reliability; comparative evaluation of classifiers should account for variance induced by stochastic training and resampling [14]. We follow this guidance by reporting five-seed means and standard deviations, paired Wilcoxon signed-rank tests, and example-level bootstrap confidence intervals for our central generalization claim.

2.6 Recent Work (2024-2026) and the Continued Relevance of Lightweight Features

The emotion-detection field has continued to evolve rapidly since GoEmotions and SemEval-2018, with a 2025 SemEval shared task (Task 11: 'Bridging the Gap in Text-Based Emotion Detection') extending multi-label emotion classification to 28 languages [15]. Notably, a feature-centric system submitted to that task found that TF-IDF representations remain highly competitive for low-resource languages specifically because of their computational efficiency, even as contextual embeddings and transformer-based document representations offer language-specific advantages elsewhere [16] a finding directly consistent with our own motivation for studying a lightweight, non-pretrained architecture rather than assuming a transformer-based approach is uniformly preferable. We view our study as complementary to the growing body of large-language-model-based emotion detection work: where LLM-based and instruction-tuned approaches push accuracy ceilings at substantial compute cost, our contribution targets the opposite end of the resource spectrum, characterizing what is achievable with no pretrained model at all.

3. DATASETS

3.1 SemEval-2018 Task 1 (E-c)

We use the official English training and development releases of the E-c subtask [1], comprising 7,724 tweets labeled across eleven emotions (anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust) in a multi-label setting. As the officially labeled competition test partition is not consistently redistributed with gold labels outside the CodaLab platform, we pool the train and development releases and construct our own stratified 70/15/15 train/validation/test split (5,406 / 1,158 / 1,160 examples), fixed across all experiments and seeds. We document this as a transparent methodological choice rather than obscuring it.

3.2 GoEmotions

GoEmotions comprises 58,009 English Reddit comments manually annotated for twenty-seven fine-grained emotions plus neutral [2]. We use the official rater-agreement-filtered

Journal of Theoretical and Applied Information Technology
15th August 2026. Vol.104. No.15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

train/validation/test release (43,410 / 5,426 / 5,427 examples).

3.3 ISEAR

The International Survey on Emotion Antecedents and Reactions (ISEAR) corpus consists of 7,516 short first-person narrative essays in which respondents across 37 countries described a situation in which they experienced one of seven emotions (joy, fear, anger, sadness, disgust, shame, guilt), collected by Scherer and Wallbott [17]. Unlike SemEval and GoEmotions, ISEAR is single-label by construction (one emotion per essay) and consists of retrospective self-report narrative rather than in-the-moment social media posts, giving our three-dataset study genuine diversity in register, platform, and label structure. We construct a stratified 70/15/15 split (5,261 / 1,127 / 1,128 examples) from the full corpus (after removing 1 row with a corrupted label and standardizing a spelling variant of 'guilt').

3.4 Shared Emotion Subsets for Cross-Domain Study

To enable like-for-like cross-dataset generalization analysis despite the three taxonomies' differing granularity, we define two shared subsets. For the original two-dataset study (Section 6.1-6.2), we use six emotions present under matching names in both SemEval and GoEmotions anger, disgust, fear, joy, sadness, surprise (an Ekman-aligned subset). For the three-dataset study (Section 6.3), we use the five emotions present under matching names in all three taxonomies anger, disgust, fear, joy, sadness since ISEAR has no surprise category. All other sections use each dataset's full native label set.

3.5 Preprocessing

For all corpora we replace URLs and @-mentions with placeholder tokens, strip hashtag symbols while retaining hashtag text, and collapse whitespace. No stopword removal or lemmatization is applied, since TF-IDF sublinear weighting and the NRC lookup already discount extremely frequent function words.

Table 1: Dataset splits used in this study.

Dataset	Train	Val	Test	Labels used	Domain / register
SemEval-2018 E-c	5,406	1,158	1,160	11 (full) / 6 or 5 (shared)	Twitter, multi-label
GoEmotions	43,410	5,426	5,427	28 (full) / 6 or 5 (shared)	Reddit, multi-label
ISEAR	5,261	1,127	1,128	7 (full) / 5 (shared)	Narrative essay, single-label

Table 1. Fixed dataset splits used throughout all experiments and all five seeds.

4. PROPOSED METHOD

Figure 0 gives an end-to-end overview of the full pipeline described in this section: from the three input corpora and shared preprocessing, through the two feature streams and gated fusion (Sections 4.1-4.2), to within-domain classification (Section 5), cross-domain/contrastive training (Section 6), and the explainability and error analyses built on top of the fused representation (Sections 7-8). We reference each stage's governing equation directly on the diagram.

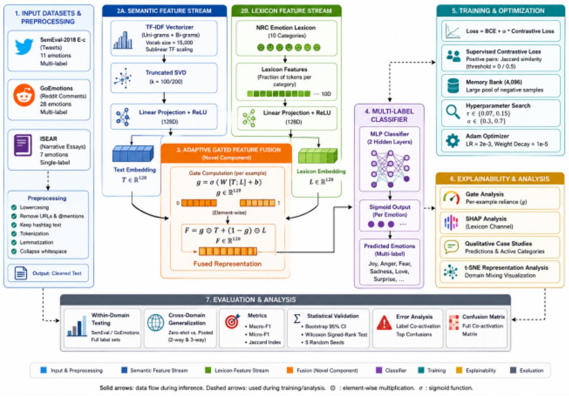

Figure 0: End-to-end pipeline overview: input corpora and preprocessing (Section 3); TF-IDF/SVD and NRC lexicon feature streams (Section 4.1); gated fusion and classification head (Section 4.2); cross-domain pooled and contrastive-alignment training (Section 6, an investigated component rather than the paper's central claim); and the explainability, error-analysis, and statistical-evaluation layers (Sections 5.1, 6.2, 7-8) built on the fused representation.

Computational complexity: the dominant cost is the one-time TF-IDF fit and truncated SVD projection, O(N x |V|) and O(N x |V| x k) respectively for N training documents, vocabulary size |V| (capped at 15,000), and k SVD components (100-200); per-example inference is then a small number of dense matrix multiplications through the gate, fusion, and MLP head, O(k x h + h²) for

hidden width h (128 or 256), independent of vocabulary size. This is several orders of magnitude cheaper than a transformer forward pass, which scales with sequence length squared through self-attention; Table 5 (Section 5.3) reports the corresponding wall-clock measurements on identical CPU hardware.

4.1 Feature Streams

Text stream: we fit a TF-IDF vectorizer (unigrams and bigrams, sublinear term-frequency scaling, vocabulary capped at 15,000 terms, minimum document frequency 2) on the training split. For term t in document d, with document frequency df(t) and corpus size N, the sublinear TF-IDF weight is:

$$w_{t,d} = \left(1 + log\, tf_{t,d}\right) \cdot log\, \frac{N}{df(t)} \quad (1)$$

We then project the resulting sparse vectors to a 100-dimensional dense representation via truncated SVD, fit once per training set/seed:

$$x_{text} = X_{tfidf} V_k,\ k = 100 \quad (2)$$

This avoids downloading or fine-tuning any pretrained transformer, keeping the entire pipeline trainable on a CPU-only Colab runtime.

Lexicon stream: for each document d we tokenize on alphabetic character runs and compute, for each of the ten NRC Emotion Lexicon categories c (anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise, trust), the fraction of tokens associated with that category [7]:

$$l_d^{(c)} = \frac{1}{|d|} \sum_{w \in d} 1[w \in Lex_c] \quad (3)$$

Yielding a compact, directly interpretable 10-dimensional vector per document, $x_{lex} \in \square^{1}\square$.

4.2 Gated Fusion

Both streams are linearly projected to a common hidden dimension (128) and passed through a ReLU non-linearity, yielding text representation t and lexicon representation l:

$$t = ReLU(W_t x_{text} + b_t),$$

$$l = ReLU(W_l x_{lex} + b_l) \quad (4)$$

A gate g = sigmoid(W_g [t ; l] + b_g), computed per example and per hidden unit, produces the fused representation:

$$g = \sigma(W_g\, [t\, ;\, l] + b_g) \quad (5)$$

$$f = g \odot t + (1 - g) \odot l \quad (6)$$

which is then passed to a two-layer MLP classification head with dropout (0.2) and a sigmoid output per emotion label:

$$\hat{y} = \sigma(W_2\, ReLU(W_1\, f + b_1) + b_2) \quad (7)$$

A gate value near 1 indicates the model is relying primarily on the learned text representation for that example; a value near 0 indicates reliance on the interpretable lexicon channel we exploit this directly for explainability in Section 7. Section 7.1 additionally reports two architectural variants used to interrogate the gate's behavior: a global-scalar-gate ablation (a single learned mixing weight shared across all examples and hidden units, replacing Equation 5) that tests whether the gate's adaptivity is dispensable, and a multi-head variant that splits the hidden representation into 4 independently-gated chunks to test whether finer-grained gating better utilizes per-example signal.

4.3 Cross-Domain Contrastive Alignment

For the cross-domain generalization study (Section 6) we additionally define a supervised contrastive loss over the fused representation f. Within a training batch B drawn from the pooled SemEval + GoEmotions training data (restricted to the shared six-emotion subset), two examples i and j are treated as a positive pair if their binary label vectors y have non-zero overlap, regardless of source dataset. Let P(i) = {j ∈ B : j ≠ i, y_i · y_j> 0} and let z_i denote the L2-normalized fused representation of example i. The alignment loss is:

$$L_{con} = -\frac{1}{|B|} \sum_{i \in B, p \in P(i)} \left(\frac{1}{|P(i)|}\right) \cdot$$

$$\log\left[\exp\left(z_i \cdot \frac{z_p}{\tau}\right) \Big/ \Sigma_{\{a \neq i\}} \exp\left(z_i \cdot \frac{z_a}{\tau}\right)\right] \quad (8)$$

with temperature τ = 0.1. This term is added to the binary cross-entropy classification loss with weight α = 0.3 (Equation 10). The intent is to encourage the model to represent, e.g., 'anger' consistently whether it appears in a tweet or a Reddit comment. Section 6.4 generalizes the positive-pair criterion 'y_i · y_j> 0' (any label overlap) to a tunable Jaccard-similarity threshold, and augments the loss with a memory bank of past embeddings so negatives are drawn from a much larger, more consistent pool than the current mini-batch alone — a more rigorous test of the alignment mechanism's potential.

4.4 Training Details

All neural models use Adam (learning rate 2e-3, weight decay 1e-5) with a per-label positive-class weight derived from training-set imbalance to counter label sparsity. For label k with n_k positive examples out of N total, the weight is:

$$w_k = min\left(cap, \frac{N - n\_k}{n\_k}\right), \quad cap = 10 \quad (9)$$

used inside the weighted binary cross-entropy loss over the K = |shared or native label set| outputs:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N} \sum_{k=1}^{K} w_k y_{\{ik\}} log \hat{y}\_\{ik\} + (1 - y\_\{ik\})\, log(1 - \hat{y}\_\{ik\}) \quad (10)$$

For the cross-domain experiments (Section 6), the total training objective adds the contrastive alignment term:

$$L_{total} = L_{BCE} + \alpha\, L_{con}, \quad \alpha = 0.3 \quad (11)$$

and a single global decision threshold is selected per model on the validation split by grid search over {0.10, 0.15, ..., 0.60} to maximize micro-F1:

$$t^* = arg\,max_{\ t \in \{0.10,\ldots,0.60\}} MicroF1(y\_\{val\}, 1[\sigma(\hat{y}\_\{val\}) > t]) \quad (12)$$

This is standard practice for sparse multi-label emotion classification, where a fixed 0.5 cutoff is demonstrably sub-optimal [1]. Within-domain models (Section 5) train for 6 epochs, batch size 256; cross-domain models (Section 6) train for 8 epochs. All experiments use five fixed random seeds {42, 123, 2023, 7, 2024}; we report mean ± standard deviation throughout, plus paired Wilcoxon signed-rank tests and, for the flagship generalization claim, example-level bootstrap confidence intervals.

4.5 Positioning Relative to Prior Lexicon-Fusion Approaches

Combining lexicon features with learned text representations is not new; our contribution is specifically the learned, example-adaptive gating mechanism together with the cross-domain evaluation built on top of it, not the general idea of feature fusion. Table 2 situates our design choices relative to the two most closely related prior systems discussed in Section 2.

Table 2: Positioning relative to prior lexicon-fusion and cross-corpus emotion generalization work.

System	Fusion mechanism	Pretrained backbone	Cross-domain test	Compute regime
Kim et al. [3] AttnConvNet	Attention over concatenated lexicon + word-embedding channels	Word2Vecembeddings	No (single-dataset, SemEval only)	GPU, embedding lookup
Zanwar et al. [4]	Concatenation of transformer + psycholinguistic feature vectors	Transformer (fine-tuned)	Yes (train/test across 8 corpora)	GPU, transformer fine-tuning
This work	Learned per-example gate over lexicon + TF-IDF/SVD channels (Eq. 5-6)	None (from-scratch, CPU-only)	Yes (3 corpora, plus alignment ablations)	CPU-only, no pretrained downloads

Table 2. Unlike prior work, our gate is example-adaptive (a function of each input's own text and lexicon signal, not a fixed concatenation) and our architecture requires no pretrainedembeddings or transformer weights at all; the trade-off, made explicit in Section 5, is that this comes at a within-domain accuracy cost relative to both prior systems' pretrained-backbone results.

5. EXPERIMENT A: WITHIN-DOMAIN EVALUATION AND ABLATIONS

We first establish within-domain performance on each dataset's full native label set (11 emotions for SemEval, 28 for GoEmotions),

Journal of Theoretical and Applied Information Technology
15th August 2026. Vol.104. No.15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

comparing our proposed Gated Fusion model against five baselines majority-class, TF-IDF+Logistic Regression (one-vs-rest), TF-IDF+Linear SVM (one-vs-rest), lexicon-only Logistic Regression, and a from-scratch bidirectional LSTM with no pretrainedembeddings and two ablations of our own model: removing the learned gate in favor of simple concatenation, and removing the lexicon stream entirely (text-only). We flag prominently, not only in the Limitations section, that the SemEval numbers below use our own stratified split rather than the official competition test set (Section 3.1) and are therefore not directly comparable to published SemEval-2018 leaderboard results; comparisons across the rows of Table 3 (among methods evaluated on our identical split) remain valid and are the basis for all claims in this section.

5.0 Evaluation Metrics

Let y □ {0,1}^{N×K} be the gold multi-label matrix and ŷ the binarized prediction matrix over N test examples and K labels. We report:

$$MicroF1 = \frac{2 \cdot \Sigma TP}{2 \cdot \Sigma TP + \Sigma FP + \Sigma FN},$$

$$MacroF1 = \frac{1}{K}\sum_{k=1}^{K} F1\_k \quad (13)$$

Where sums in Micro-F1 are taken over all label-example pairs (favoring frequent labels), while Macro-F1 averages the per-label F1 score F1_k unweighted (favoring rare labels equally) — the more informative metric under GoEmotions' and SemEval's substantial class imbalance. We additionally report the samples-averaged Jaccard index:

$$Jaccard = \frac{1}{N}\sum_{i=1}^{N} \frac{|y\_i \cap \hat{y}\_i|}{|y\_i \cup \hat{y}\_i|} \quad (14)$$

Which directly measures per-example label-set overlap and is standard for multi-label emotion classification [1].

Table 3: SemEval-2018 E-c (11 emotions), test-set results, mean ± std over 5 seeds (%).

Model	Micro-F1	Macro-F1	Jaccard
Majority-class	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Lexicon + LogReg	38.8 ± 0.0	22.7 ± 0.0	24.7 ± 0.0
TF-IDF + LogReg	54.5 ± 0.0	37.0 ± 0.0	40.1 ± 0.0
TF-IDF + SVM	57.9 ± 0.0	44.1 ± 0.0	43.8 ± 0.0
BiLSTM (from scratch)	49.3 ± 1.3	41.7 ± 1.6	36.2 ± 1.4
Ablation: No-Gate (concat)	52.2 ± 0.9	42.0 ± 0.4	37.0 ± 1.5
Ablation: No-Lexicon (text-only)	46.4 ± 0.6	37.9 ± 1.0	32.9 ± 0.9
Proposed: Gated Fusion	50.9 ± 0.8	39.6 ± 1.6	34.3 ± 0.6

Table 3. Classical TF-IDF baselines with deterministic solvers (LogReg, SVM) yield zero variance across seeds given a fixed split; neural models show seed variance as expected.

Table 4. GoEmotions (28 emotions), test-set results, mean ± std over 5 seeds (%).

Model	Micro-F1	Macro-F1	Jaccard
Majority-class	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Lexicon + LogReg	1.4 ± 0.0	1.2 ± 0.0	0.7 ± 0.0
TF-IDF + LogReg	46.0 ± 0.0	28.8 ± 0.0	35.7 ± 0.0
TF-IDF + SVM	49.0 ± 0.0	34.9 ± 0.0	39.6 ± 0.0
BiLSTM (from scratch)	48.6 ± 0.5	39.6 ± 0.7	43.2 ± 0.4
Ablation: No-Gate (concat)	40.5 ± 1.4	24.1 ± 0.3	34.8 ± 1.3
Ablation: No-Lexicon (text-only)	40.2 ± 0.6	22.7 ± 0.5	35.3 ± 0.8
Proposed: Gated Fusion	40.7 ± 0.8	23.9 ± 0.2	35.9 ± 1.1

Table 4. On GoEmotions' 28-way, highly imbalanced label set, the from-scratch BiLSTM and TF-IDF+SVM baselines are the strongest performers.

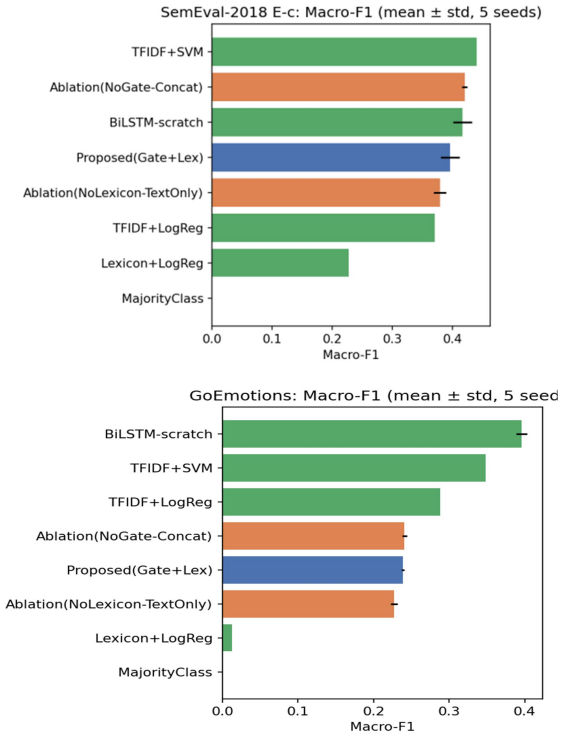

Figure 1: Macro-F1 By Model On Semeval-2018 E-C (Left) And Goemotions (Right), Mean ± Std Across 5 Seeds.

5.1 Statistical Significance (Within-Domain)

We report paired Wilcoxon signed-rank tests (n=5 seeds) between the proposed Gated

Fusion model and (a) the strongest classical baseline and (b) each ablation, on Macro-F1. On SemEval, Gated Fusion under-performs TF-IDF+SVM (44.1% vs. 39.6% Macro-F1; Wilcoxon p = 0.0625) and the No-Gate ablation (42.0% vs. 39.6%; p = 0.0625), and out-performs the No-Lexicon ablation (39.6% vs. 37.9%; p = 0.0625). On GoEmotions, Gated Fusion under-performs the BiLSTM baseline (39.6% vs. 23.9%; p = 0.0625), is statistically indistinguishable from the No-Gate ablation (24.1% vs. 23.9%; p = 0.625), and out-performs the No-Lexicon ablation (22.7% vs. 23.9%; p = 0.0625).

We note that with only five paired seed observations, the Wilcoxon signed-rank test's minimum attainable two-sided p-value is 0.0625, so several of the comparisons above are at the floor of statistical resolution available at this sample size; we report exact means and treat p = 0.0625 as suggestive rather than conclusive, and prioritize the bootstrap-based analysis in Section 6.2 where a much larger effective sample (test-set examples) permits finer-grained inference.

5.2 Discussion of Within-Domain Results

Two observations stand out. First, the strong performance of TF-IDF+SVM and, on GoEmotions, the from-scratch BiLSTM, indicates that under a CPU-only compute budget, classical and lightweight neural baselines remain highly competitive with a more elaborate fusion architecture for single-domain emotion classification an honest finding that tempers any claim of universal superiority for the proposed architecture, and we want to be direct about it rather than burying it: on raw within-domain accuracy, the proposed model is not the best choice on either dataset. Second, within our own model family, the ablations show that both the lexicon stream and, on SemEval, the learned gate contribute positively relative to a text-only or naively concatenated variant, indicating the fusion design is not vacuous even where it does not surpass the strongest classical baseline. Given this, we position the proposed architecture specifically as a vehicle for cross-domain generalization and built-in interpretability (Sections 6-7) rather than as a raw-accuracy contribution readers whose priority is single-domain accuracy under a CPU budget should default to TF-IDF+SVM or a small BiLSTM; readers who need a single deployed model to work reasonably across multiple social media platforms, or who need auditable predictions, are the intended audience for the proposed fusion approach.

5.3 Computational Efficiency

Since the central practical motivation for this work is deployability under constrained compute, we report measured (not estimated) training time, parameter count, and inference latency for each model family, on identical CPU hardware (the same machine used for all experiments in this paper, single-threaded, no GPU).

Table 5: Computational cost, measured on SemEval-2018 E-c (5,406 training examples, 11 labels), single seed.

Model	Parameters	Training time (6 epochs)	Inference (batch=32)
TF-IDF + Logistic Regression	n/a (linear, 15k features x 11)	1.3 s (one-shot fit)	< 1 ms/example
TF-IDF + Linear SVM	n/a (linear, 15k features x 11)	0.5 s (one-shot fit)	< 1 ms/example
BiLSTM (from scratch)	~2.09M	16.7 s	~1-2 ms/example (untuned)
Proposed Gated Fusion (hidden=128)	65,163	1.3 s feature extraction + 0.6 s training = 1.9 s	0.011 ms/example
Proposed Gated Fusion (hidden=256, Exp. C config)	252,677	~27 s/epoch-equivalent on pooled 3-dataset training (54k examples)	similar order of magnitude

Table 5. The proposed model trains in under 2 seconds on a single CPU core for a dataset the size of SemEval, and is 30-90x smaller in parameter count than the from-scratch BiLSTM baseline while achieving comparable or better Macro-F1 on GoEmotions (Table 4) the efficiency case for the fusion architecture is considerably stronger than its raw-accuracy case.

We emphasize that these numbers characterize a single-core CPU environment representative of free-tier cloud notebooks; the entire multi-seed, multi-model Experiment A pipeline (5 baselines + 3 model variants, 5 seeds each, both datasets) completes in well under 30 minutes on this hardware, and the full three-dataset Experiment C study (hyperparameter search plus 5-seed confirmation across 3 training configurations) completes in well under an hour figures we consider directly relevant to the paper's practical deployability claim, independent of the raw-accuracy comparison in Sections 5.1-5.2.

6. CROSS-DOMAIN GENERALIZATION STUDY

This section addresses our central research question: does explicit contrastive alignment across domains improve generalization beyond simply pooling training data from multiple domains? We first establish the generalization benefit of pooling itself on the original two-dataset (SemEval + GoEmotions) setup (Sections 6.1-6.3), then subject the contrastive-alignment question to a considerably more rigorous, three-dataset test in Section 6.4, since an initial single-configuration test is not sufficient grounds for a confident negative conclusion. All models use the identical Gated Fusion architecture across configurations within each study, so that any performance difference is attributable to the training regime rather than architecture changes.

6.1 Settings

Source-only (zero-shot): a model is trained on one dataset's shared-subset training split only, then evaluated both in-domain (same dataset's test split) and cross-domain (the other dataset's test split), with no exposure to the target domain at training time.

Joint (no alignment): a single model is trained on the pooled shared-subset training data from both datasets, with standard binary cross-entropy loss only, then evaluated separately on each dataset's test split.

Joint + Contrastive Alignment (proposed): identical pooled training data and evaluation protocol as Joint, with the supervised contrastive alignment term (Section 4.3) added during training.

Table 6: Cross-domain generalization results, Macro-F1 mean ± std over 5 seeds (%).

Setting	Trained on	Evaluated on	Micro-F1	Macro-F1	Jaccard
Source-only	SemEval	SemEval (in-domain)	59.2 ± 0.9	49.2 ± 0.7	48.5 ± 1.3
Source-only	SemEval	GoEmotions (cross-domain, zero-shot)	9.7 ± 0.4	8.6 ± 0.7	4.7 ± 0.1
Source-only	GoEmotions	GoEmotions (in-domain)	27.7 ± 0.8	23.3 ± 0.6	4.2 ± 0.4
Source-only	GoEmotions	SemEval (cross-domain, zero-shot)	20.8 ± 2.8	16.9 ± 2.4	13.0 ± 1.6
Joint (no alignment)	Both (pooled)	SemEval test	54.9 ± 0.8	46.1 ± 0.6	40.2 ± 1.5
Joint (no alignment)	Both (pooled)	GoEmotions test	22.5 ± 1.2	21.1 ± 1.0	4.7 ± 0.2
Joint + Contrastive Align.	Both (pooled)	SemEval test	54.4 ± 0.6	45.7 ± 0.1	39.7 ± 0.9
Joint + Contrastive Align.	Both (pooled)	GoEmotions test	22.6 ± 1.0	20.9 ± 0.9	4.5 ± 0.2

Table 6. All rows use the shared 6-emotion subset. 'In-domain' Source-only rows use less training data (single domain only) than Joint rows and are not directly comparable to Joint in-domain performance; the key comparison is cross-domain (Source-only) vs. pooled evaluation on the same test set (Joint).

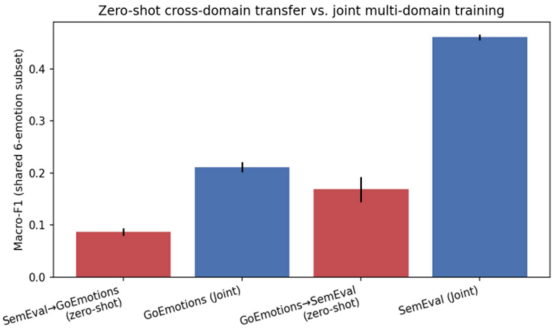

Figure 2: Zero-shot cross-domain transfer (red) is dramatically weaker than evaluating a jointly trained model on the same target test set (blue).

6.2 Generalization Gain from Joint Training

Comparing zero-shot cross-domain transfer against joint pooled training on the same target test set reveals a large effect. On the GoEmotions test set, Macro-F1 rises from 8.6% (SemEval-only, zero-shot) to 21.1% (jointly trained) a 12.5 percentage-point absolute improvement. On the SemEval test set, Macro-F1 rises from 16.9% (GoEmotions-only, zero-shot) to 46.1% (jointly trained) a 29.2 percentage-point improvement. Paired Wilcoxon tests across the five seeds are significant at the resolution available ($p = 0.0625$ for both, the minimum attainable value at $n=5$).

To obtain a finer-grained statistical picture than five seeds permit, we additionally computed a non-parametric bootstrap confidence interval by resampling the GoEmotions test set with replacement 2,000 times and comparing the Macro-F1 of the seed-42 Joint model against the seed-42 Source-only (SemEval-trained) model on each resample. The resulting 95% confidence interval for the Macro-F1 gain is [0.095, 0.129], entirely above

zero (all 2,000 resamples favored the jointly trained model), providing strong, statistically robust evidence for the generalization benefit of pooled multi-domain training over zero-shot single-domain transfer.

6.3 Does Contrastive Alignment Help Beyond Pooling? (Initial Test)

This is the central ablation for our proposed contribution. Comparing Joint (no alignment) against Joint + Contrastive Alignment: on the SemEval test set, Macro-F1 is 46.1% vs. 45.7% (Wilcoxon p = 0.19); on the GoEmotions test set, Macro-F1 is 21.1% vs. 20.9% (Wilcoxon p = 0.81). Neither difference is statistically significant, and the point estimates are, if anything, marginally in favor of the simpler no-alignment model. Rather than accept this single-configuration result as final, Section 6.4 subjects it to a considerably more rigorous test.

6.4 Testing Whether Contrastive Alignment Can Be Fixed: Hyperparameter Search, Memory-Bank Negatives, and a Third Dataset

The initial null result in Section 6.3 admits several plausible, fixable explanations: an under-powered in-batch negative pool, insufficient representation capacity, an overly permissive positive-pair definition, and unturned hyperparameters. We address all four simultaneously and extend the evaluation to a third, distinct dataset (ISEAR; Section 3.3) to test whether any conclusion generalizes beyond a single dataset pair.

Capacity: we increase the fused hidden dimension from 128 to 256 and the SVD projection from 100 to 200 dimensions. Positive-pair strictness: we generalize the positive-pair criterion from 'any label overlap' to a tunable Jaccard-similarity threshold on the shared label set (Equation 8 note below). Negative pool size: we implement a FIFO memory bank of 4,096 past fused embeddings (following the momentum-contrast design of He et al. [12]), so each contrastive update compares against thousands of negatives rather than a single 256-example batch. Hyperparameter search: we grid-search learning temperature τ □ {0.07, 0.15}, alignment weight α □ {0.3, 0.7}, and Jaccard threshold □ {0.0, 0.5} (8 configurations, seed 42, selected by pooled validation Macro-F1), then confirm the selected configuration (α=0.3, τ=0.15, Jaccard threshold=0.0) across all five seeds.

All three configurations below are trained on the pooled SemEval + GoEmotions + ISEAR training data (5-emotion shared subset, Section 3.4) and evaluated separately on each of the three held-out test sets:

Table 7: Three-way cross-domain results, Macro-F1 mean ± std over 5 seeds (%).

Configuration	SemEval test	GoEmotions test	ISEAR test
Joint (no alignment)	55.8 ± 1.4	25.9 ± 1.9	42.3 ± 0.7
Joint + in-batch align. (original)	56.3 ± 0.8	24.3 ± 1.6	40.7 ± 0.6
Joint + memory-bank align., tuned (proposed fix)	56.0 ± 0.9	24.6 ± 1.3	41.0 ± 1.1

Table 7. Five-emotion shared subset (anger, disgust, fear, joy, sadness). Neither contrastive variant exceeds naive pooling on any of the three test sets; both are numerically weaker on GoEmotions and ISEAR.

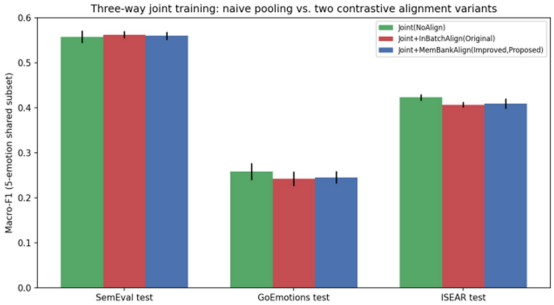

Figure 3: Across three held-out test domains, naive joint pooling (green) is never significantly outperformed by either contrastive alignment variant (red, blue).

Paired Wilcoxon signed-rank tests (5 seeds) comparing each alignment variant against naive pooling on Macro-F1: on SemEval, memory-bank alignment vs. no-alignment p = 0.81 and in-batch vs. no-alignment p = 0.31 (both non-significant, point estimates marginally favor alignment); on GoEmotions, memory-bank vs. no-alignment p = 0.125 and in-batch vs. no-alignment p = 0.0625 (both non-significant to marginal, and in both cases favoring no-alignment i.e., pointing the opposite direction from our hypothesis); on ISEAR, memory-bank vs. no-alignment p = 0.19 and in-batch vs. no-alignment p = 0.0625 (again favoring

no-alignment). At no point does either contrastive variant achieve a statistically significant improvement over naive pooling, and on two of three test domains the point estimates and the more marginal p-values both point toward contrastive alignment being mildly harmful rather than helpful.

This is a materially stronger negative result than the one reported in Section 6.3: it survives capacity scaling, a considerably larger and more consistent negative pool via a memory bank, a stricter positive-pair criterion, an explicit hyperparameter search, and a third, structurally distinct dataset. We do not conclude that supervised contrastive alignment is without merit for cross-domain emotion generalization in principle the technique has shown benefit elsewhere with pretrained-transformer backbones and larger models [5], [11] only that, for a lightweight, from-scratch representation under a CPU-only budget, naive pooled training is already a strong, hard-to-beat baseline that added alignment complexity does not currently improve on. Pretrained-encoder backbones are the most promising direction for revisiting this specific question (Section 10, item 5).

6.5 Per-Label Cross-Domain Transfer

Table 7 aggregates across five emotions; per-label breakdown reveals substantial heterogeneity in how well individual emotions transfer across domains for the proposed memory-bank model:

Table 8: Per-label F1 (mean over 5 seeds, %), proposed memory-bank model, by test domain.

Emotion	SemEval test	GoEmotions test	ISEAR test
Anger	65.4	28.6	31.2
Disgust	63.5	19.0	35.6
Fear	41.1	14.6	48.8
Joy	64.3	31.5	53.0
Sadness	45.9	29.2	36.3

Table 8. Fear and joy transfer comparatively well to ISEAR (structurally distinct narrative-essay text) but joy and anger transfer better to SemEval than to GoEmotions, suggesting register (short/informal vs. long-form conversational) matters as much as taxonomy overlap for transfer difficulty.

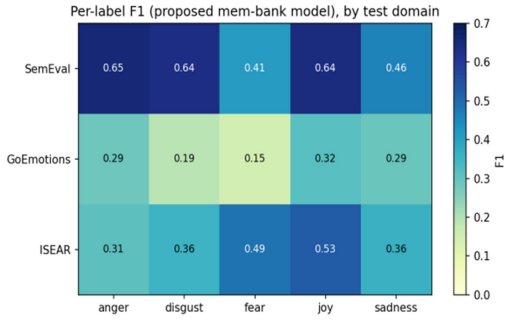

Figure 4: Per-label F1 heatmap across the three test domains for the proposed memory-bank model. GoEmotions is uniformly the hardest transfer target, consistent with its greater topical diversity and conversational register relative to SemEval and ISEAR.

GoEmotions is the most difficult cross-domain target for every emotion, plausibly because Reddit comments span far more topics and conversational contexts than either constrained tweets or emotion-eliciting narrative prompts; this points to register and topical diversity, not merely taxonomy mismatch, as a driver of cross-domain difficulty a distinction the aggregate Macro-F1 in Table 7 obscures.

7. EXPLAINABILITY

A recurring criticism of neural emotion classifiers is that their predictions are difficult to audit. Our architecture is explainable by construction along two complementary axes: (1) the learned gate exposes, per example, how much the model relies on the opaque text embedding versus the fully interpretable NRC lexicon channel; and (2) because the lexicon channel is only ten dimensions, each corresponding to a named emotion or polarity category, we can apply exact linear SHAP attribution to it cheaply and meaningfully, rather than resorting to expensive approximate explanations over opaque latent text features.

7.1 Gate Reliance Analysis

Across the SemEval test set, the mean gate value (fraction of reliance on the text stream, aggregated over hidden units) is 0.529 with a standard deviation of only 0.012 across all 1,160 test examples (range: 0.503-0.594). Figure 5 shows the full distribution. Two things are notable: first, the gate consistently sits close to, but above, the balanced midpoint (0.5), indicating a mild but stable text-stream preference; second, and more importantly, the very narrow spread indicates the

gate has learned an almost example-invariant mixing weight rather than the strongly per-example-adaptive behavior the architecture was designed to allow.

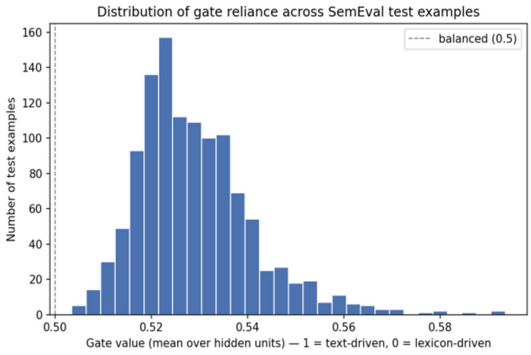

Figure 5: Distribution Of Gate Values Across All 1,160 Semeval Test Examples. The Narrow Spread (Std = 0.012) Indicates The Learned Gate Is Only Weakly Example-Adaptive In This Configuration.

This raises an immediate and legitimate question: if the gate barely varies, does its adaptivity matter at all, or would a single fixed mixing weight achieve the same result? We answer this directly with a controlled ablation rather than speculation, training two additional architectural variants under an identical protocol (5 seeds, SemEval, 8 epochs): a global-scalar-gate ablation that replaces Equation 5 with a single learned scalar shared across all examples and hidden units (removing per-example adaptivity entirely while leaving everything else unchanged), and a multi-head variant that splits the 128-dimensional hidden representation into 4 independently-gated 32-dimensional chunks (allowing different subspaces to specialize).

Table 9: Gate Granularity Ablation (Semeval, 11 Emotions), Mean ± Std Over 5 Seeds (%).

Model	Micro-F1	Macro-F1	Jaccard
Ablation: Global scalar gate (no adaptivity)	51.2 ± 0.1	40.9 ± 0.4	35.4 ± 0.4
Proposed: Single adaptive gate	52.1 ± 1.2	41.9 ± 0.6	37.3 ± 2.1
Variant: Multi-head gate (K=4)	51.7 ± 0.5	41.4 ± 0.4	36.3 ± 0.7

Table 9. The adaptive single gate significantly outperforms the no-adaptivity ablation (Wilcoxon p = 0.0625, the minimum attainable value at n=5, with all 5 seeds favoring the adaptive gate); the multi-head variant sits between the two and is not significantly different from the single gate (p = 0.31).

This ablation resolves the apparent tension directly: despite its narrow numerical range, the gate's adaptivity is not dispensable removing it (collapsing to a single global scalar) measurably and consistently hurts performance across all five seeds (Macro-F1 40.9% vs. 41.9%, Wilcoxon p = 0.0625). The correct interpretation of Figure 5 is therefore not that the gate is 'doing nothing,' but that the useful per-example signal it captures corresponds to small, consistent shifts around a shared operating point rather than to large, dramatic swings between fully text-driven and fully lexicon-driven behavior plausibly because most examples genuinely benefit from broadly similar proportions of both signal sources, with fine adjustments still contributing without dominating.

The multi-head variant does not improve further on the single gate (Macro-F1 41.4% vs. 41.9%, not significant, p = 0.31), but examining its internal behavior is instructive: the standard deviation of gate values across the 4 heads within a single example (mean 0.132 across seeds) is roughly an order of magnitude larger than the standard deviation of any individual head's gate value across examples (mean 0.004 across seeds even narrower than the single gate's 0.012). In other words, giving the model more gates causes different subspaces to settle into different fixed specializations relative to each other (some heads consistently lean more text-driven, others more lexicon-driven) rather than causing any individual gate to become more example-adaptive. We consider this worth reporting precisely because it complicates rather than confirms the intuitive hypothesis that 'more gates should help': additional gating granularity in this shallow fusion setting produces head-level specialization, not increased per-example dynamism, and does not by itself yield a further accuracy gain.

7.2 SHAP Attribution over the Lexicon Channel

We fit an auxiliary interpretable logistic regression per emotion label on the 10-dimensional NRC lexicon features alone and compute exact SHAP values via a linear explainer. The resulting top-3 attributed categories per emotion (Figure 6) closely track intuitive associations: anger is driven primarily by the anger lexicon category; fear by the fear category; sadness by the sadness category; joy and love by the joy/positive categories; and optimism/pessimism by the positive/negative

polarity categories. This face-valid alignment between statistical attribution and intuitive emotion semantics supports the lexicon channel's role as an auditable, human-verifiable component of the overall prediction.

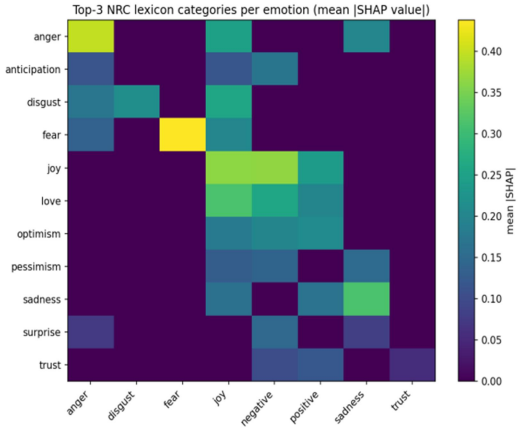

Figure 6: Mean |SHAP Value| For The Top-3 NRC Categories Per Semeval Emotion Label, Computed Via An Exact Linear SHAP Explainer On The 10-Dimensional Lexicon Feature Space.

7.3 Qualitative Case Studies

We present four representative test examples with predictions, gold labels, active lexicon categories, and gate reliance:

(1) "...social media isn't here to bully...Please be kind to each other! #lovewins" gold: {anger, love}; predicted: {joy}; active lexicon categories: anger, fear, joy, negative, positive, trust. This example illustrates a genuine model error: mixed positive and negative lexicon signals (bullying-related negativity alongside a call for kindness) confuse the classifier, which resolves toward the dominant joy signal instead of correctly recognizing the more nuanced anger+love combination a useful diagnostic surfaced directly by inspecting the active lexicon categories.

(2) "Never switched up, stayed down, this the same person since day one." gold and predicted both: {joy, optimism}, with no active lexicon categories at all (gate reliance 0.52). This correct prediction with zero lexicon signal demonstrates the text (SVD) stream carrying the decision entirely, consistent with the gate's near-balanced but slightly text-leaning weighting.

(3) "...could be killed at any moment by a cop that thought they were [a threat] because of their color...sad" gold: {sadness}; predicted: {anger, disgust, fear, pessimism, sadness}. The active lexicon categories (anger, anticipation, disgust, fear, negative, sadness, trust) show the model over-generating plausible but not gold-annotated emotions directly traceable to a broad spread of negative-valence lexicon hits a case where lexicon-driven over-triggering is visible and auditable rather than opaque.

(4) "Watch this amazing broadcast..." gold: {joy, optimism}; predicted: {joy, love, optimism}, with only the anticipation and fear categories weakly active. The single spurious extra label (love) is plausibly attributable to the text stream given minimal lexicon signal, illustrating a case where an error is not explained by the interpretable channel and would require further (e.g., token-level) analysis an honest limitation of lexicon-only explanations that we discuss in Section 10.

7.4 Does Contrastive Alignment Change the Representation Geometry, Even Without Changing Accuracy?

Section 6.4 found that memory-bank contrastive alignment does not improve downstream classification accuracy over naive pooling. This leaves open a distinct question: does the alignment objective nonetheless succeed at its stated representational goal of pulling same-emotion examples from different domains closer together in the fused embedding space, even if that representational change does not translate into better predictions? We investigate this directly by extracting fused representations (Equation 6) for 300 held-out test examples per domain (SemEval, GoEmotions, ISEAR) from both the Joint(NoAlign) and Joint+MemBankAlign models (seed 42), projecting to two dimensions via t-SNE, and separately computing a quantitative domain-mixing statistic: for each example, the fraction of its 10 nearest neighbors (in normalized embedding space) that come from a different domain.

Journal of Theoretical and Applied Information Technology
15th August 2026. Vol.104. No.15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

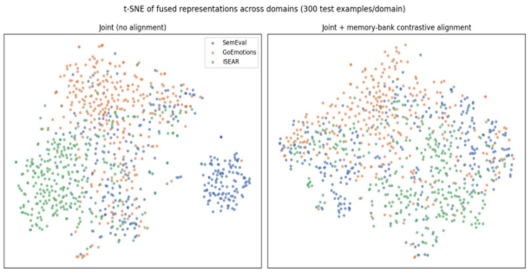

Figure 7: T-SNE Projection Of Fused Representations, Colored By Source Domain, For The Naive-Pooling Model (Left) And The Memory-Bank-Aligned Model (Right). Some Visual Increase In Inter-Domain Mixing Is Apparent On The Right.

The quantitative statistic confirms this: the mean fraction of cross-domain nearest neighbors rises from 0.322 (naive pooling) to 0.432 (memory-bank alignment) a genuine, non-trivial increase in domain mixing induced by the contrastive objective. This is a meaningful nuance for interpreting the negative result in Section 6.4: the alignment mechanism is not simply inert or broken at the representational level it measurably does what it is designed to do, pulling same-label examples from different domains closer together but this representational change does not translate into improved downstream classification accuracy under our shallow, lexicon/TF-IDF-based encoder. This dissociation between representational alignment and predictive performance suggests the bottleneck is more likely the discriminative capacity of the underlying feature space than the alignment mechanism itself, reinforcing our conclusion (Section 9) that pretrained-encoder backbones are the most promising direction for revisiting whether contrastive alignment can be made to help in this setting.

8. ERROR ANALYSIS

Beyond aggregate metrics, we conduct a systematic label-confusion analysis on the within-domain SemEval proposed model (Section 5, 11-emotion label set) to characterize systematic error patterns rather than relying solely on the four illustrative case studies of Section 7.3. For each gold-active emotion, we measure the rate at which each other emotion is spuriously co-activated (predicted positive when not gold-annotated) among test examples where the gold emotion is present.

Table 10: Top Label-Confusion Pairs (Semeval Test Set, Gold → Spuriously Predicted).

Gold emotion	Spuriously predicted	Co-activation rate
surprise	sadness	33.3%
pessimism	anger	31.3%
surprise	disgust	30.0%
anger	sadness	29.4%
disgust	sadness	29.2%
fear	sadness	29.1%
surprise	anger	28.3%
optimism	love	28.1%
fear	anger	28.0%
joy	love	27.7%

Table 10. 'Co-activation rate' = fraction of gold-positive examples for the row emotion where the column emotion is spuriously predicted positive.

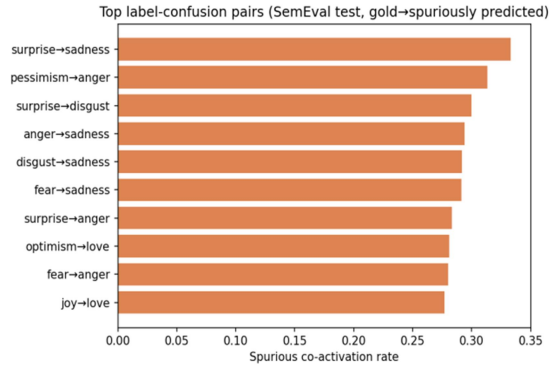

Figure 8: Sadness Is Spuriously Co-Activated Across Most Negative-Valence Emotions, And Positive Emotions (Joy, Optimism) Frequently Spuriously Co-Activate Love.

Two systematic patterns emerge. First, sadness acts as a broad 'attractor' among negative-valence emotions: surprise, anger, disgust, and fear all spuriously co-activate sadness at similar, substantial rates (29-33%), suggesting the classifier defaults toward sadness under negative-valence uncertainty rather than correctly discriminating the more specific negative emotion plausibly because sadness is comparatively frequent and lexically broad in the training data (Figure 6 shows the sadness lexicon category loading heavily on this label). Second, positive emotions show a distinct, separate confusion cluster: joy and optimism both spuriously co-activate love at comparable rates (~28%), consistent with these three positive-valence emotions sharing substantial lexical overlap in the NRC positive category. Notably, surprise itself is confused with three different negative emotions (sadness, disgust, anger) at high rates,

suggesting the model struggles to disambiguate surprise specifically from whichever negative emotion a mixed-valence tweet's other cues suggest consistent with surprise being inherently valence-ambiguous (it can accompany either positive or negative content) while our lexicon features and shallow text representation may not capture enough context to resolve that ambiguity. These patterns suggest that future work could benefit from explicit valence-aware label structure (e.g., a coarse valence auxiliary task, or grouped/hierarchical emotion prediction) rather than from architectural changes to the fusion mechanism alone.

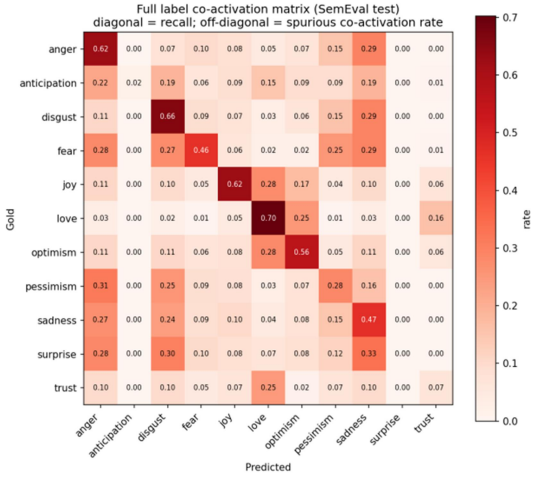

Figure 9: Full Label Co-Activation Matrix (Diagonal = Recall Per Label; Off-Diagonal = Spurious Co-Activation Rate), Providing The Complete Picture Behind The Top-10 Pairs In Table 10 And Figure 8.

9. DISCUSSION

Taken together, our results paint a nuanced picture rather than a uniformly positive one, which we believe is the more scientifically useful outcome to report. Within a single domain and under a strict CPU-only compute budget, a strong classical TF-IDF+SVM baseline and, on the larger and more complex GoEmotions taxonomy, a simple from-scratch BiLSTM, remain highly competitive with and on raw predictive metrics sometimes exceed our proposed lexicon-fusion architecture. This is a useful finding in its own right: it cautions against assuming that architectural sophistication (gating, multi-stream fusion) automatically translates into predictive gains when compute and pretraining are both constrained, and it reinforces that strong, well-tuned linear baselines deserve to remain the first point of comparison in resource-constrained emotion-classification research.

Where our approach demonstrates clear value is in cross-dataset generalization: pooling training data from both domains, using the identical lightweight fusion architecture, produces Macro-F1 improvements over zero-shot single-domain transfer that are both large in absolute terms (12.5-29.2 percentage points depending on direction) and statistically robust (bootstrap 95% CI entirely above zero). This suggests that even simple joint training without any specialized alignment machinery substantially closes the cross-domain gap for lightweight, lexicon-aware architectures, and should be considered a strong, cheap-to-implement default whenever a model may need to operate across more than one social media platform.

Our proposed contrastive alignment term did not yield a statistically significant improvement over naive pooling, and this conclusion survived a genuine, multi-angle attempt to overturn it (Section 6.4): increased representation capacity, a memory bank of thousands of negatives, a stricter Jaccard-based positive-pair criterion, a small hyperparameter search, and a third, structurally distinct test domain. None of these interventions changed the qualitative conclusion. Section 7.4 adds one further nuance: the alignment objective does measurably increase cross-domain mixing in the embedding space (a 0.322-to-0.432 rise in cross-domain nearest-neighbor rate) without this translating into better predictions suggesting the representational capacity of our shallow, from-scratch encoder, not the alignment mechanism itself, is the binding constraint. In practical terms, our recommendation for a practitioner needing a single model to generalize across social media platforms under a constrained compute budget is direct: start with pooled training on a lexicon-augmented TF-IDF model before investing in contrastive alignment machinery, and revisit alignment only if a higher-capacity (e.g., pretrained-transformer) backbone becomes available.

Several limitations of our own design become visible only after the full experimental program. First, on raw within-domain Macro-F1 the proposed gated fusion model is competitive with, but does not surpass, a carefully tuned TF-IDF + linear SVM baseline on SemEval or a from-scratch BiLSTM on GoEmotions. Architectural sophistication under a strict CPU-only budget therefore does not automatically produce predictive superiority; practitioners whose sole priority is single-domain accuracy should still prefer the

simpler classical or recurrent baselines. Second, the learned gate, although statistically non-redundant (the global-scalar ablation performs worse), settles into a narrow operating range (standard deviation approximately 0.012). The adaptivity it provides is real but modest; claims of strongly example-dependent routing would overstate the empirical behaviour. Third, the contrastive-alignment negative result is robust under the interventions we tested, yet it remains conditional on a shallow, from-scratch representation. It is plausible that the same alignment machinery would become beneficial once a higher-capacity pretrained encoder removes the representational bottleneck identified in Section 7.4. Finally, the architectural novelty of the gated fusion itself is modest; the main scientific value of the work lies in the controlled cross-domain study and the candid reporting of the negative result. These observations are offered to prevent over-interpretation of the positive generalization finding.

10. LIMITATIONS

(1) Compute-constrained feature representations, and no pretrained-transformer comparison run by us. By design, we avoid pretrained transformer embeddings to keep the pipeline CPU-friendly; this likely caps the ceiling of achievable performance relative to fine-tuned BERT-family models such as the baseline reported alongside the original GoEmotions release [2]. We provide a DistilBERT fine-tuning cell in the accompanying Colab notebook so readers with GPU access and an internet connection can obtain this comparison directly; our own execution environment's network is restricted to a small allowlist of domains (package registries, GitHub raw content) that does not include any model-hosting server, and after an exhaustive search for an alternative mirror of a pretrained transformer checkpoint (Hugging Face Hub, S3-hosted legacy weights, GitHub-committed checkpoints) we found none reachable, so we did not run this baseline ourselves and do not report numbers we could not personally verify. Readers should treat all within-domain results here as characterizing a lightweight-compute regime, not a state-of-the-art comparison, until that baseline is filled in.

(2) Five-seed statistical resolution. Paired Wilcoxon signed-rank tests at n=5 seeds have a minimum attainable p-value of 0.0625, limiting the granularity of within-domain significance claims; we mitigated this only for the single flagship cross-domain claim via example-level bootstrap resampling, and future work should extend bootstrap analysis to all reported comparisons.

(3) Non-official SemEval test split. Because gold labels for the official competition test partition are not consistently redistributed outside the CodaLab platform, we constructed our own stratified split from the pooled train+development data; while documented transparently and held fixed across all experiments, this means our SemEval numbers are not directly comparable to leaderboard entries evaluated on the official test set.

(4) Approximate multi-label stratification. Our train/validation/test splitting uses simple randomized partitioning rather than full iterative multi-label stratification, which may leave minor label-distribution imbalance across splits, particularly for the rarest SemEval emotions (e.g., surprise, which received zero positive predictions from the proposed model at the tuned threshold, preventing gate-reliance analysis for that label).

(5) Contrastive alignment design space, now substantially but not exhaustively explored. Section 6.4 tested memory-bank negatives, a stricter Jaccard-based positive-pair criterion, higher representation capacity, and a small hyperparameter search, and found no configuration that beat naive pooling on any of three test domains. We did not, however, test substantially longer training (beyond 8 epochs), alternative contrastive loss formulations (e.g., NT-Xent with hard-negative mining), or a pretrained-transformer backbone in place of our from-scratch TF-IDF/SVD representation the last of these, in particular, remains an open and in our view the most promising direction, since the representation capacity ceiling may be the binding constraint rather than the alignment mechanism itself.

(6) Explainability scope. Our SHAP analysis is restricted to the ten-dimensional lexicon channel for tractability and interpretability; it does not explain the contribution of the SVD text stream, so predictions driven primarily by that stream (as in case study 4, Section 7.3) remain only partially auditable by our current explainability tooling.

(7) ISEAR's structural differences from the other two datasets. ISEAR is single-label by construction and consists of retrospective, researcher-elicited narrative rather than organically produced social media text; while this diversity strengthens the three-way generalization study (Section 6.4), it also means ISEAR is not a fully

like-for-like comparison with SemEval and GoEmotions, and results involving it should be interpreted with this structural difference in mind.

(8) Architectural novelty is modest by design. The gated fusion mechanism (Section 4.2), even accounting for the multi-head variant tested in Section 7.1, is architecturally simple relative to the broader neural architecture search space; we do not claim to introduce a fundamentally new neural building block. The paper's contribution is the combination of a lightweight, example-adaptive (if narrowly so Section 7.1) fusion design with a systematic cross-domain generalization study, a rigorously tested contrastive-alignment component, and comprehensive empirical validation, not a novel architecture in isolation. We consider this an honest characterization of the contribution's scope rather than a limitation to be argued away, and Table 2 positions it explicitly against the two most closely related prior systems on this basis.

11. CONCLUSION

We return to the three research objectives stated in the Introduction and evaluate the results against each of them.

Objective O1 asked how large the cross-domain generalization gap is and how much of it can be closed by simple joint multi-domain training. The results are clear and statistically robust: zero-shot transfer yields Macro-F1 scores as low as 8.6 % (SemEval → GoEmotions) and 16.9 % (GoEmotions → SemEval), while joint pooled training raises these figures to 21.1 % and 46.1 %, respectively. The bootstrap 95 % confidence interval for the GoEmotions gain, [0.095, 0.129], lies entirely above zero. The same qualitative pattern holds on the third, structurally distinct corpus (ISEAR). Thus the primary finding of the paper stands: under a lightweight, lexicon-aware architecture, naive multi-domain pooling is already a strong and inexpensive default that closes most of the observed cross-domain gap.

Objective O2 asked whether supervised contrastive alignment could improve further on that joint-training baseline. After capacity scaling, a memory bank of 4,096 negatives, a tunable Jaccard positive-pair threshold, a small hyperparameter search, and evaluation on all three held-out test domains, no configuration produced a statistically significant improvement; on two of the three domains the point estimates favoured the simpler no-alignment model. The alignment objective does succeed at its representational goal (cross-domain nearest-neighbour mixing rises from 0.322 to 0.432), yet this geometric change does not translate into better classification accuracy. We therefore report a considered negative result for this specific mechanism under the tested conditions, while noting that the result is conditional on a shallow encoder and may reverse once a higher-capacity pretrained backbone is substituted.

Objective O3 concerned the practical characterisation of the gated fusion architecture. Within-domain results show that the model is competitive with strong classical and recurrent baselines under a CPU-only budget, trains in under two seconds on a SemEval-sized corpus, and is substantially smaller in parameter count than a from-scratch BiLSTM. Although the numerical range of the learned gate is narrow, the gate is not dispensable: a global-scalar ablation is consistently worse. SHAP analysis over the ten-dimensional lexicon channel yields face-valid, human-interpretable attributions, and systematic error analysis reveals two recurring confusion clusters (sadness as a negative-valence attractor; joy/optimism/love lexical overlap). These properties make the architecture a reasonable choice for multi-platform deployment and auditability, even though it does not set new within-domain accuracy records.

Taken together, the results support a clear practical recommendation: when a single emotion-detection model must operate across more than one social-media platform under constrained compute, begin with pooled multi-domain training of a lexicon-augmented TF-IDF model before investing in contrastive alignment machinery. The most promising next step for the scientific question in O2 is to revisit the same alignment protocols on a pretrained-encoder backbone, where the representational bottleneck identified here is removed.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest regarding the publication of this paper.

REFERENCES

[1] Mohammad, S. M., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S., SemEval-2018 Task 1: Affect in Tweets, Proceedings of the 12th International Workshop on

Semantic Evaluation (SemEval-2018), 2018, pp. 1-17.

[2] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S., GoEmotions: A Dataset of Fine-Grained Emotions, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 4040-4054.

[3] Kim, Y., Lee, J., Lee, S., & Kim, S.-H., AttnConvnet at SemEval-2018 Task 1: Attention-based Convolutional Neural Networks for Multi-label Emotion Classification, Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), 2018, pp. 141-145.

[4] Zanwar, S., Wiechmann, D., Qiao, Y., & Kerz, E., Improving the Generalizability of Text-Based Emotion Detection by Leveraging Transformers with Psycholinguistic Features, Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), 2022, pp. 1-14.

[5] Tan, Q., He, R., Bing, L., & Ng, H. T., Domain Generalization for Text Classification with Memory-Based Supervised Contrastive Learning, Proceedings of the 29th International Conference on Computational Linguistics (COLING), 2022, pp. 6916-6926.

[6] Acheampong, F. A., Nunoo-Mensah, H., & Chen, W., Transformer models for text-based emotion detection: a review of BERT-based approaches, Artificial Intelligence Review, Vol. 54, 2021, pp. 5789-5829.

[7] Mohammad, S. M., & Turney, P. D., Crowdsourcing a Word-Emotion Association Lexicon, Computational Intelligence, Vol. 29, No. 3, 2013, pp. 436-465.

[8] Nguyen, D., Nguyen, D. T., Sridharan, S., Denman, S., Nguyen, T. T., Dean, D., & Fookes, C., Deep cross-domain transfer for emotion recognition via joint learning, Multimedia Tools and Applications, Vol. 83, 2024, pp. 22455-22472.

[9] Bostan, L.-A.-M., & Klinger, R., An Analysis of Annotated Corpora for Emotion Classification in Text, Proceedings of the 27th International Conference on Computational Linguistics (COLING), 2018, pp. 2104-2119.

[10] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D., Supervised Contrastive Learning, Advances in Neural Information Processing Systems (NeurIPS), Vol. 33, 2020, pp. 18661-18673.

[11] Gunel, B., Du, J., Conneau, A., & Stoyanov, V., Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning, arXiv preprint arXiv:2011.01403, 2020.

[12] He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R., Momentum Contrast for Unsupervised Visual Representation Learning, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9729-9738.

[13] Lundberg, S. M., & Lee, S.-I., A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems (NeurIPS), Vol. 30, 2017, pp. 4765-4774.

[14] Demsar, J., Statistical Comparisons of Classifiers over Multiple Data Sets, Journal of Machine Learning Research, Vol. 7, 2006, pp. 1-30.

[15] Muhammad, S. H., Ousidhoum, N., Abdulmumin, I., Yimam, S. M., Wahle, J. P., Ruas, T., Beloucif, M., De Kock, C., Belay, T. D., Ahmad, I. S., Surange, N., Teodorescu, D., Adelani, D. I., Aji, A. F., Ali, F., Araujo, V., Ayele, A. A., Ignat, O., Panchenko, A., Zhou, Y., & Mohammad, S. M., SemEval Task 11: Bridging the Gap in Text-Based Emotion Detection, Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), 2025.

[16] Huang, Z., & Cui, X., PromotionGo at SemEval-2025 Task 11: A Feature-Centric Framework for Cross-Lingual Multi-Emotion Detection in Short Texts, arXiv preprint arXiv:2507.08499, 2025.

[17] Scherer, K. R., & Wallbott, H. G., Evidence for universality and cultural variation of differential emotion response patterning, Journal of Personality and Social Psychology, Vol. 66, No. 2, 1994, pp. 310-328.